

Convergence of cyclic coordinatewise ℓ_1 minimization

Kshitij Khare and Bala Rajaratnam

University of Florida and Stanford University

Abstract

We consider the general problem of minimizing an objective function which is the sum of a convex function (not strictly convex) and absolute values of a subset of variables (or equivalently the ℓ_1 -norm of the variables). This problem appears extensively in modern statistical applications associated with high-dimensional data or “big data”, and corresponds to optimizing ℓ_1 -regularized likelihoods in the context of model selection. In such applications, cyclic coordinatewise minimization (CCM), where the objective function is sequentially minimized with respect to each individual coordinate, is often employed as it offers a computationally cheap and effective optimization method. Consequently, it is crucial to obtain theoretical guarantees of convergence for the sequence of iterates produced by the cyclic coordinatewise minimization in this setting. Moreover, as the objective corresponds to flat ℓ_1 -regularized likelihoods of many variables, it is important to obtain convergence of the iterates themselves, and not just the function values. Previous results in the literature only establish either, (i) that every limit point of the sequence of iterates is a stationary point of the objective function, or (ii) establish convergence under special assumptions, or (iii) establish convergence for a different minimization approach (which uses quadratic approximation based gradient descent followed by an inexact line search), (iv) establish convergence of only the function values of the sequence of iterates produced by random coordinatewise minimization (a variant of CCM). In this paper, a rigorous general proof of convergence for the cyclic coordinatewise minimization algorithm is provided. We demonstrate the usefulness of our general results in contemporary applications by employing them to prove convergence of two algorithms commonly used in high-dimensional covariance estimation and logistic regression.

1 Introduction

Let $g : \mathbb{R}^m \rightarrow \mathbb{R} \cup \{\infty\}$ be a twice differentiable strictly convex function, whose effective domain has a non-empty interior C_g . Suppose also that g has a positive curvature everywhere on C_g , and that $g(\mathbf{t})$ converges to infinity as \mathbf{t} approaches the boundary of C_g . Let S be a given subset of $\{1, 2, \dots, n\}$, E be an $m \times n$ matrix having no zero column, and $\lambda > 0$ be fixed. Let

$$\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n : x_i \geq 0 \text{ for every } i \in S^c\}.$$

Define the functions f_1 and f_2 , where $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$ for $i = 1, 2$ as follows:

$$f_1(\mathbf{x}) = g(E\mathbf{x}) + \lambda \sum_{i \in S} |x_i|,$$

and,

$$f_2(\mathbf{x}) = \mathbf{x}^T E^T E \mathbf{x} - \sum_{i \in S^c} \log x_i + \lambda \sum_{i \in S} |x_i|.$$

Consider the following two minimization problems:

$$\text{Minimize } f_1(\mathbf{x}) \text{ subject to } \mathbf{x} \in \mathcal{X}. \quad (1.1)$$

$$\text{Minimize } f_2(\mathbf{x}) \text{ subject to } \mathbf{x} \in \mathcal{X}. \quad (1.2)$$

The minimization problems in (1.1) and (1.2) appear extensively in contemporary applications, and are particularly relevant in statistics and machine learning (see for example [9, 12, 10, 15, 22, 28, 31, 32]), and signal processing (see for example [3, 5, 6, 7, 29, 30]). In statistical applications, the function $g(E\mathbf{x})$ is typically a log-likelihood or pseudo log-likelihood corresponding to a statistical model. Traditional statistical methods focus on minimizing the function $g(E\mathbf{x})$ without the addition of a minimizer like $\|\mathbf{x}\|_1$. However, in the modern context of high-dimensional data or “big data”, it is often desirable to obtain sparse solutions (solutions where many entries are exactly equal to zero), resulting in the inclusion of the term $\lambda \sum_{i \in S} |x_i|$ in the objective function. The most challenging features of the minimization problems (1.1) and (1.2) are the following:

1. In many applications, $m < n$. Hence the functions $g(E\mathbf{x})$, $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are not strictly convex, and in general do not have a unique global minimum.
2. The minimization occurs on a high-dimensional space, with hundreds or thousands (if not more) of variables.
3. The minimization problem is non-smooth due to the presence of the “ ℓ_1 penalty” term $\lambda \sum_{i \in S} |x_i|$.

Hence, any method proposed for finding a solution to the above problem should be computationally scalable and have theoretical convergence guarantees. In many statistical applications involving high dimensional regression and high dimensional covariance estimation (see for example [10, 13, 15]), coordinatewise minimization can be performed in closed form. Hence, for such problems, a cyclic coordinatewise minimization (CCM) algorithm (where each iteration consists of minimizing the objective function sequentially over all the coordinates) is often used, as it offers a computationally cheap and effective method for minimizing the respective objective functions. In situations where coordinatewise minimization cannot be achieved in closed form, it often involves minimizing a one-dimensional convex function, and can be numerically achieved to a high degree of accuracy in a few steps. Hence, coordinatewise minimization has also been used in such situations (see for example [28]). Hence, understanding the convergence properties of the cyclic coordinatewise minimization algorithm for (1.1) and (1.2) is a crucial and relevant task. However, a rigorous proof of convergence of the cyclic coordinatewise minimization algorithm for minimization problems in (1.1) and (1.2) is not available in the literature. We now provide a brief overview of existing optimization methods and convergence results related to these problems.

In Tseng [33], it is proved that under appropriate conditions on g , every limit point of the sequence of iterates produced by the cyclic coordinatewise minimization algorithm is

a stationary point of the corresponding objective function. However, this does not necessarily mean that the sequence of iterates converges. Tseng and Yun [34] propose a block-coordinatewise gradient descent (CGD) approach, which can be thought of as a hybrid of gradient-projection and coordinate descent. In particular, they consider minimizing an objective function of the form

$$f(\mathbf{x}) + c \sum_{i=1}^n P_j(x_j), \quad (1.3)$$

where P_j is a proper, convex, lower semicontinuous function for $1 \leq j \leq n$, and f is a continuously differentiable function on an open subset of \mathbb{R}^n containing the effective domain of P_j for every $1 \leq j \leq n$. At each iteration, a quadratic approximation of the function f is considered, a descent direction is then generated by applying block coordinate descent, followed by an inexact line search along this direction (by using an Armijo-type rule to ensure sufficient descent). The authors in [34] also provide a proof that the sequence of iterates produced by their algorithm converges under suitable assumptions. Note that if the function g in (1.1) is not quadratic, then clearly the CCM and CGD approaches are distinctly different. Considering (1.2), we note that it can be expressed in the framework of (1.3) in two ways. We can choose $f(\mathbf{x}) = \mathbf{x}^T E^T E \mathbf{x} - \sum_{i \in S^c} \log x_i$ and $P_j(x_j) = |x_j|$ for $j \in S$, in which case the CCM and CGD approaches are again different as f is not quadratic. Alternatively, if we choose $f(\mathbf{x}) = \mathbf{x}^T E^T E \mathbf{x}$, $P_j(x_j) = |x_j|$ for $j \in S$, and $P_j(x_j) = -\log x_j$ for $j \in S^c$, then the function $\sum_{j=1}^n P_j(x_j)$ is not polyhedral. Hence, the assumptions in [34, Lemma 7] do not apply, and it is not clear if the convergence results in [34, Theorem 2] and [34, Theorem 3] apply.

Saha and Tewari [27] provide finite time convergence results for a variety of cyclic coordinatewise descent methods for objective functions of the form $f(\mathbf{x}) + \lambda \sum_{j=1}^n |x_j|$. However, their convergence results rely on the assumption that the function f is isotone, i.e., essentially f is twice-differentiable and the Hessian matrix of f at any \mathbf{x} in its effective domain has non-positive off-diagonal entries. Such an assumption does not hold in general for many contemporary applications and those which we consider in Section 5.

Luo and Tseng [20] consider the following minimization problem.

$$\text{Minimize } g(E\mathbf{x}) + \mathbf{b}^T \mathbf{x} \text{ subject to } l_i \leq x_i \leq u_i \forall 1 \leq i \leq n, \quad (1.4)$$

where $\mathbf{b}, \mathbf{l}, \mathbf{u}$ are fixed n -dimensional vectors. The entries of \mathbf{l} and \mathbf{u} are allowed to be $-\infty$ and ∞ respectively. They provide a very detailed and intricate proof of convergence of the sequence of iterates produced by the cyclic coordinatewise descent algorithm for (1.4) (see also [19, 21]). Note once more that the minimization problem in (1.4) is substantially different than the minimization problem in (1.1) and (1.2).

In recent useful work, Richtarik and Takac [25] provide a random coordinatewise descent algorithm for solving (1.3), where instead of cycling over all the coordinate blocks, a random coordinate is chosen and minimized over at each iteration. Intuitively speaking, a randomized choice of coordinates may avoid a possible worst case ordering of the coordinates in the cyclic setting, is also more suitable for situations when all the data is not available all the time, and more amenable for a convergence analysis. The authors in [25] establish important convergence (and provides rates) for the function values of the sequence of iterates produced by the random coordinatewise descent algorithm. Establishing convergence of the sequence

of *iterates* for random coordinatewise descent however remains a challenge. We note that one of the compelling reasons that has motivated the use of random coordinatewise descent algorithm versus the (non-random) cyclic coordinatewise minimization is that the former allows for easier convergence analysis, though many methods that have been proposed in the machine learning and statistics literature actually use the (non-random) cyclic coordinatewise minimization. In this paper we address the crucial and challenging problem of establishing convergence of the sequence of iterates in the (non-random) cyclic coordinatewise minimization setting. We also note that in modern high-dimensional problems in Statistics/Machine Learning, the objective functions are often very flat, and it is quite likely that although the function values converge, the sequence of iterates do not.

Several methods other than cyclic coordinatewise minimization have also been proposed in the literature to solve the minimization problems in (1.1) and (1.2) (or a more general version of this problem, where the term $\lambda \sum_{i \in S} |x_i|$ is replaced by a (block) separable non-smooth function). One class of methods is based on proximal gradient descent with an Armijo-type stepsize (see for example [11, 16]). Another class of methods is based on trust-regions (see for example [1, 4, 8]). See Tseng and Yun [34] for a detailed list of related references. We note that none of these methods correspond to the classical coordinatewise minimization approach that has been proposed and extensively used in the statistical applications outlined above.

In this paper, we provide a rigorous proof of convergence of the sequence of iterates produced by the cyclic coordinatewise minimization algorithm for the minimization problems in (1.1) and (1.2). We shall build on the work of Luo and Tseng [20], and extend it when incorporating non-differentiable terms of the form $\lambda \sum_{i \in S} |x_i|$ in $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$. This generalization makes the convergence analysis of the cyclic coordinatewise minimization algorithm for (1.1) and (1.2) more complex as compared to the convergence analysis of the cyclic coordinatewise minimization approach for (1.4). We shall see that the non-smooth term leads to many challenging and non-trivial questions.

This paper is organized as follows. In Section 2, we provide a summary of the assumptions, algorithms and the main convergence results in the paper. A detailed proof of convergence for the cyclic coordinatewise minimization algorithm for the minimization problems in (1.1) and (1.2) is then provided in Section 3 and Section 4 respectively. The results in Section 3 and Section 4 are then used in Section 5 to establish convergence of two algorithms arising in high-dimensional covariance estimation and high dimensional logistic regression.

2 Summary of main results

In this section, we undertake the following: (a) provide the assumptions that are made for the minimization problems in (1.1) and (1.2), (b) formally define the cyclic coordinatewise minimization algorithms corresponding to these problems, and (c) state the main convergence results that are established later in this paper. Recall that $f_1(\mathbf{x}) = g(E\mathbf{x}) + \lambda \sum_{i \in S} |x_i|$. We start by providing the assumptions that will be made for the minimization problem in (1.1).

- (A1) The effective domain of g has a non-empty interior C_g .
- (A2) g is strictly convex and twice continuously differentiable on C_g .

- (A3) Either $g(t) \rightarrow \infty$ as t approaches the boundary of C_g , or, $|S| = n$ and g is non-negative with $C_g = \mathbb{R}^m$.
- (A4) g has a positive curvature everywhere on C_g .
- (A5) The set of optimal solutions of the minimization problem in (1.1), denoted by \mathcal{X}^* , is non-empty.

Consider the following practical implementation of the coordinatewise descent (CCM) algorithm to solve the minimization problem in (1.1).

Algorithm 1 Cyclic coordinatewise descent algorithm for f_1

1. Set $r = 0$. Start with initial value $\mathbf{x}^0 \in \mathcal{X}$ such that $f_1(\mathbf{x}^0)$ is finite, and a prespecified tolerance ϵ .
2. Set $\mathbf{x}^{r,0} = \mathbf{x}^r$.
3. For $i = 1, 2, \dots, n$, set

$$\mathbf{x}^{r,i} = \operatorname{argmin}_{\mathbf{x} \in \mathcal{X}: x_j = x_j^{r,i-1} \forall j \neq i} f_1(\mathbf{x}). \quad (2.1)$$

4. Set $\mathbf{x}^{r+1} = \mathbf{x}^{r,n}$. If $\|\mathbf{x}^{r+1} - \mathbf{x}^r\| > \epsilon$, set $r = r + 1$, return to Step 2. Otherwise, stop.
-

We first claim that (2.1) is well-defined by using contradiction. Note that for any $\xi \in \mathbb{R}$, the set $H_\xi := \{E\mathbf{x} : \mathbf{x} \in \mathcal{X}, f_1(\mathbf{x}) \leq \xi\}$ is contained in the set $\{E\mathbf{x} : \mathbf{x} \in \mathcal{X}, g(E\mathbf{x}) \leq \xi\}$. It follows by [20, Lemma A.1] that if $g(t) \rightarrow \infty$ as t approaches the boundary of C_g , then $\{E\mathbf{x} : \mathbf{x} \in \mathcal{X}, g(E\mathbf{x}) \leq \xi\}$ is bounded. Alternatively, if $|S| = n$ and g is non-negative, then H_ξ is contained in the set $\{E\mathbf{x} : \sum_{i=1}^n |x_i| \leq \xi/\lambda\}$. In either case, we get that

$$H_\xi \text{ is bounded for every } \xi \in \mathbb{R}. \quad (2.2)$$

Suppose now that the minimum in (2.1) is not attained for some r and i . Let \mathbf{e}^i denote the i^{th} unit vector in \mathbb{R}^n . There are then two possibilities:

- (a) $i \in S$ and $f_1(\mathbf{x}^{r,i-1} - h\mathbf{e}^i)$ is non-increasing as $h \rightarrow \infty$. Hence, $\mathbf{x}^{r,i-1} - h\mathbf{e}^i \in H_{f_1(\mathbf{x}^{r,i-1})}$ for large enough h . The boundedness of $H_{f_1(\mathbf{x}^{r,i-1})}$ implies that $E\mathbf{e}^i = 0$, which contradicts the assumption that no column of E is zero.
- (b) $f_1(\mathbf{x}^{r,i-1} + h\mathbf{e}^i)$ is non-increasing as $h \rightarrow \infty$. This case leads to the same contradiction as in (a).

The following theorem now formally establishes convergence of the sequence of iterates produced by Algorithm 1, and is the first main result in this paper.

Theorem 2.1 *The sequence of iterates $\{\mathbf{x}^r\}_{r \geq 0}$ generated by the cyclic coordinatewise descent algorithm for f_1 converges to a value $\mathbf{x}^* \in \mathcal{X}$ such that $f_1(\mathbf{x}^*) \leq f_1(\mathbf{x})$ for every $\mathbf{x} \in \mathcal{X}$.*

A detailed proof of Theorem 2.1 will be provided in Section 3. We now briefly outline the major steps in the proof. We first show that the difference between the successive iterates produced by the CCM algorithm for f_1 goes to zero. Further arguments establish that this sequence of differences between the successive iterates is actually square-summable. Note that square-summability of the sequence of differences is not sufficient to establish that $\{\mathbf{x}^r\}_{r \geq 0}$ is a Cauchy sequence. We then proceed to show that the distance between the sequence of iterates and the boundary of \mathcal{X}^* (the set of optimal solutions) goes to zero. Again, this itself is also not sufficient to establish that $\{\mathbf{x}^r\}_{r \geq 0}$ is a Cauchy sequence (see the discussion just before Lemma 3.7). However, using the three facts above, along with some matrix-theoretic results and combinatorial arguments, we prove that the sequence of iterates produced by the cyclic coordinatewise descent algorithm is a Cauchy sequence with limit $\mathbf{x}^* \in \mathcal{X}$. This is done as follows. First, we show that eventually some coordinates of \mathbf{x}^r are exactly equal to zero, while the remaining coordinates are bounded from zero as $r \rightarrow \infty$. Second, we show that (see Lemma 3.12) the coordinates of \mathbf{x}^r that stay away from zero are influenced by those coordinates which eventually become zero. Moreover, this influence is a function of the distance between these ultimate zero coordinates and zero, and therefore dies away as $r \rightarrow \infty$. This is then used, along with a series of combinatorial arguments, to establish that (see Lemma 3.14) for an arbitrary $\epsilon > 0$, there exists an $\mathbf{x}^* \in \mathcal{X}^*$ such that $\|\mathbf{x}^r - \mathbf{x}^*\| < \epsilon$ for large enough r . This immediately implies that $\{\mathbf{x}^r\}_{r \geq 0}$ is a Cauchy sequence. Since we have already established that the distance between the sequence of iterates and the boundary of \mathcal{X}^* goes to zero, convergence to an optimal solution follows.

Now, we consider the problem of minimizing the function f_2 defined in (1.2). Recall that

$$\mathbf{x}^T E^T E \mathbf{x} - \sum_{i \in S^c} \log x_i + \lambda \sum_{i \in S} |x_i|.$$

For the function f_2 , the only assumption that is made is a stronger version of assumption (A5), this assumption essentially states that the level sets of f_2 are bounded, and is standard in many contemporary applications.

- (A5)* Let $\xi \in \mathbb{R}$ be arbitrarily fixed. If $\mathbf{x} \in \mathcal{X}$ satisfies $f_2(\mathbf{x}) \leq \xi$, then there exists $\xi^* \in \mathbb{R}_+$ (independent of \mathbf{x}) such that $1/\xi^* \leq x_i \leq \xi^*$ for every $i \in S^c$ and $|x_i| \leq \xi^*$ for every $i \in S$.

We shall also show that this level set assumption will be satisfied in the application considered in Section 5. Again, we consider the following coordinatewise descent algorithm to solve the minimization problem in (1.2). Note that the steps of the following algorithm are identical to that of Algorithm 1, except that f_1 is replaced by f_2 . However, we have provided separate statements of the two algorithms for expositional convenience, in particular, for differentiating between the sequence of iterates produced by applying the CCM algorithm for f_1 and f_2 .

It will be shown in Section 4 (see Lemma 4.1) that the minimization in (2.3) is well-defined, and the unique minimizer can be obtained in closed form. The following theorem establishes convergence of the sequence of iterates produced by the cyclic coordinatewise descent algorithm for minimizing f_2 and is the second main result in this paper.

Algorithm 2 Cyclic coordinatewise descent algorithm for f_2

1. Set $r = 0$. Start with initial value $\mathbf{z}^0 \in \mathcal{X}$ such that $f_2(\mathbf{z}^0)$ is finite, and a prespecified tolerance ϵ .
2. Set $\mathbf{z}^{r,0} = \mathbf{z}^r$.
3. For $i = 1, 2, \dots, n$, set

$$\mathbf{z}^{r,i} = \operatorname{argmin}_{\mathbf{x} \in \mathcal{X}, x_j = z_j^{r,i-1} \forall j \neq i} f_2(\mathbf{x}). \quad (2.3)$$

4. Set $\mathbf{z}^{r+1} = \mathbf{z}^{r,n}$. If $\|\mathbf{z}^{r+1} - \mathbf{z}^r\| > \epsilon$, set $r = r + 1$, go to Step 2.
-

Theorem 2.2 *The sequence of iterates $\{\mathbf{z}^r\}_{r \geq 0}$ generated by the cyclic coordinatewise descent algorithm for minimizing f_2 converges to a $\mathbf{z}^* \in \mathcal{X}$ such that $f_2(\mathbf{z}^*) \leq f_2(\mathbf{x})$ for every $\mathbf{x} \in \mathcal{X}$.*

A detailed proof of Theorem 2.2 will be provided in Section 4. There are two differences between the functions f_1 and f_2 . Let $q(\mathbf{y}) = \mathbf{y}^T \mathbf{y}$ for every $\mathbf{y} \in \mathbb{R}^m$. The term $g(E\mathbf{x})$ in f_1 is replaced by the special choice $q(E\mathbf{x}) = \mathbf{x}^T E^T E \mathbf{x}$ in f_2 . The presence of the logarithmic terms in f_2 however introduces a new feature as compared to f_1 . Hence, although the basic method of proving convergence remains the same for f_2 , the presence of the logarithmic terms in f_2 create new challenges which will be tackled in the convergence analysis in Section 4.

3 Convergence analysis for cyclic coordinatewise minimization applied to f_1

In this section, we provide a detailed proof of Theorem 2.1. We start with the following lemma about \mathcal{X}^* , the set of optimal solutions of the minimization problem (1.1). The proof of this lemma follows immediately from arguments in [20, Page 5] and is therefore omitted.

Lemma 3.1 *\mathcal{X}^* is a convex set. Also, there exists $\mathbf{t}^* \in \mathbb{R}^m$ such that*

$$E\mathbf{x}^* = \mathbf{t}^*, \forall \mathbf{x}^* \in \mathcal{X}^*.$$

It follows from assumptions (A1) and (A4) that $\nabla^2 g$ is positive definite in some open ball U^* containing \mathbf{t}^* . Hence, there exists $\sigma > 0$ such that

$$\nabla^2 g(\mathbf{t}) - \sigma I_n \text{ is positive definite } \forall \mathbf{t} \in U^*. \quad (3.1)$$

Let $\mathbf{d}(\mathbf{x}) = \nabla\{g(E\mathbf{x})\} = E^T \nabla g(E\mathbf{x})$, where $\nabla g(E\mathbf{x})$ denotes the gradient function of g evaluated at $E\mathbf{x}$. We denote the i^{th} entry of $\mathbf{d}(\mathbf{x})$ by $d_i(\mathbf{x})$. Let $\mathbf{d}^* := E^T \nabla g(\mathbf{t}^*)$. It follows by Lemma 3.1 that

$$\mathbf{d}(\mathbf{x}^*) = \mathbf{d}^* \forall \mathbf{x}^* \in \mathcal{X}^*. \quad (3.2)$$

Note that the sub differential versions of the KKT conditions for the convex minimization problem in (1.1) imply that $\mathbf{x} \in \mathcal{X}^*$ if and only if

$$x_i = \max(0, x_i - d_i(\mathbf{x})) \text{ for } i \in S^c, \quad (3.3)$$

$$d_i(\mathbf{x}) + \lambda \text{sign}(x_i) = 0 \text{ if } x_i \neq 0, i \in S, \quad (3.4)$$

$$|d_i(\mathbf{x})| \leq \lambda \text{ if } x_i = 0, i \in S. \quad (3.5)$$

We provide an alternative characterization of the elements of \mathcal{X}^* , which will be useful in our analysis.

Lemma 3.2 $\mathbf{x} \in \mathcal{X}^*$ if and only if

$$x_i = \max(0, x_i - d_i(\mathbf{x})) \text{ for } i \in S^c, \quad (3.6)$$

$$x_i = \text{sign}(x_i - d_i(\mathbf{x})) \max(|x_i - d_i(\mathbf{x})| - \lambda, 0) \text{ for } i \in S. \quad (3.7)$$

The proof of this lemma is provided in the appendix. Recall that $\{\mathbf{x}^r\}_{r \geq 0}$ is the sequence of iterates generated by the coordinatewise descent algorithm for minimizing f_1 , and $\mathbf{x}^{r,i}$ is the appropriate coordinatewise minimizer defined in (2.1). It follows from the arguments in the proof of Lemma 3.2 that for $i \in S$,

$$x_i^{r,i} = \text{sign}(x_i^{r,i} - d_i(\mathbf{x}^{r,i})) \max(|x_i^{r,i} - d_i(\mathbf{x}^{r,i})| - \lambda, 0), \quad (3.8)$$

and for $i \in S^c$

$$x_i^{r,i} = \max(0, x_i^{r,i} - d_i(\mathbf{x}^{r,i})). \quad (3.9)$$

Next, we state a lemma from [14] which was used in [20], and will also play an important role in our analysis. Let $\|\mathbf{x}\| := \sqrt{\mathbf{x}^T \mathbf{x}}$ denote the Euclidean norm, and $\mathbf{x}^+ := (\max(0, x_i))_{i=1}^n$ for any vector \mathbf{x} . Also, $\mathbf{x} \leq \mathbf{y}$ implies that $x_i \leq y_i$ for every $1 \leq i \leq n$.

Lemma 3.3 ([14]) *Let B_1 and B_2 be any $k_1 \times n$ and $k_2 \times n$ matrices respectively. Then, there exists a constant $\theta > 0$ depending only on B_1 and B_2 such that, for any $\bar{\mathbf{x}} \in \mathcal{X}$ and any k_1 -vector \mathbf{d}_1 and k_2 -vector \mathbf{d}_2 such that the linear system $B_1 \mathbf{y} = \mathbf{d}_1, B_2 \mathbf{y} \leq \mathbf{d}_2, \mathbf{y} \in \mathcal{X}$ is consistent, there is a point $\bar{\mathbf{y}}$ satisfying $B_1 \bar{\mathbf{y}} = \mathbf{d}_1, B_2 \bar{\mathbf{y}} = \mathbf{d}_2, \bar{\mathbf{y}} \in \mathcal{X}$, with*

$$\|\bar{\mathbf{x}} - \bar{\mathbf{y}}\| \leq \theta(\|B_1 \bar{\mathbf{x}} - \mathbf{d}_1\| + \|(B_2 \bar{\mathbf{x}} - \mathbf{d}_2)^+\|).$$

Now let

$$\mathbf{t}^{r,i} = E \mathbf{x}^{r,i}$$

for all r and all $0 \leq i \leq n$. By (2.1), it follows that

$$f_1(\mathbf{x}^{r,i}) \leq f_1(\mathbf{x}^{r,i-1}) \quad (3.10)$$

for every r and $1 \leq i \leq n$. Hence, $\mathbf{t}^{r,i} \in H_{f_1}(\mathbf{x}^0)$ for every r and $0 \leq i \leq n$. It follows by (2.2) that

$$\{\mathbf{t}^{r,i}\}_{r \geq 0, 0 \leq i \leq n} \text{ is bounded.} \quad (3.11)$$

Also, since g is twice continuously differentiable, it follows that $\{g(\mathbf{t}^{r,i})\}_{r \geq 0}$ is uniformly bounded above for all $0 \leq i \leq n$. If $g(t) \rightarrow \infty$ as t approaches the boundary of C_g , it follows

that every limit point of $\{\mathbf{t}^{r,i}\}_{r \geq 0}$ lies in C_g for all $0 \leq i \leq n$. If $C_g = \mathbb{R}^m$, then it follows by (3.11) that again every limit point of $\{\mathbf{t}^{r,i}\}_{r \geq 0}$ lies in C_g for all $0 \leq i \leq n$. By (3.10), the sequence $\{f_1(\mathbf{x}^{r,i})\}_{r \geq 0}$ decreases to the same quantity, say f^∞ for every $0 \leq i \leq n$. If $f^\infty = -\infty$, then assumption (A5) (which says that the set of optimal solutions to (1.1) is non-empty) will be violated. Hence $f^\infty > -\infty$. We now prove that the difference between the successive iterates of the cyclic coordinatewise descent algorithm for f_1 converges to zero.

Lemma 3.4

$$\|\mathbf{x}^{r+1} - \mathbf{x}^r\| \rightarrow 0 \text{ as } r \rightarrow \infty.$$

Proof We proceed by contradiction. Suppose the result does not hold. Then there exists $\epsilon > 0$, $i \in \{1, 2, \dots, n\}$ and a subsequence \mathcal{R} of \mathbb{N} such that $|x_i^{r+1} - x_i^r| > \epsilon$ for every $r \in \mathcal{R}$. It follows by the definition of $\mathbf{t}^{r,i}$ that

$$\|\mathbf{t}^{r,i} - \mathbf{t}^{r,i-1}\| = \|E(\mathbf{x}^{r,i} - \mathbf{x}^{r,i-1})\| = \|E_{\cdot i}\| |x_i^{r+1} - x_i^r| \geq \|E_{\cdot i}\| \epsilon, \quad (3.12)$$

where $E_{\cdot i}$ denotes the i^{th} column of E . Since $\{\mathbf{t}^{r,i}\}_{r \in \mathcal{R}}$ and $\{\mathbf{t}^{r,i-1}\}_{r \in \mathcal{R}}$ are bounded, we assume without loss of generality that there is a further subsequence \mathcal{R}' of \mathcal{R} such that $\{\mathbf{t}^{r,i}\}_{r \in \mathcal{R}'}$ and $\{\mathbf{t}^{r,i-1}\}_{r \in \mathcal{R}'}$ converge to \mathbf{t}' and \mathbf{t}'' respectively. It follows by (3.12) that $\mathbf{t}' \neq \mathbf{t}''$. Since $\mathbf{t}', \mathbf{t}'' \in C_g$, it follows by the continuity of g that

$$\{g(\mathbf{t}^{r,i})\}_{r \in \mathcal{R}'} \rightarrow g(\mathbf{t}'), \quad \{g(\mathbf{t}^{r,i-1})\}_{r \in \mathcal{R}'} \rightarrow g(\mathbf{t}''). \quad (3.13)$$

It follows by the definition of f that

$$\left\{ \sum_{j \in S} |x_j^{r,i}| \right\}_{r \in \mathcal{R}'} \rightarrow f^\infty - g(\mathbf{t}'), \quad \left\{ \sum_{j \in S} |x_j^{r,i-1}| \right\}_{r \in \mathcal{R}'} \rightarrow f^\infty - g(\mathbf{t}''). \quad (3.14)$$

Since $\mathbf{x}^{r,i}$ is obtained from $\mathbf{x}^{r,i-1}$ by minimizing along the i^{th} coordinate, the convexity of f yields

$$\begin{aligned} f(\mathbf{x}^{r,i}) &\leq f\left(\frac{\mathbf{x}^{r,i} + \mathbf{x}^{r,i-1}}{2}\right) = g\left(\frac{\mathbf{t}^{r,i} + \mathbf{t}^{r,i-1}}{2}\right) + \frac{\sum_{j \in S} |x_j^{r,i} + x_j^{r,i-1}|}{2} \\ &\leq g\left(\frac{\mathbf{t}^{r,i} + \mathbf{t}^{r,i-1}}{2}\right) + \frac{\sum_{j \in S} |x_j^{r,i}| + |x_j^{r,i-1}|}{2}, \end{aligned} \quad (3.15)$$

for every $r \in \mathcal{R}'$. Using the continuity of g , (3.14) and passing to the limit as $r \rightarrow \infty, r \in \mathcal{R}'$, we obtain

$$f^\infty \leq f^\infty + g\left(\frac{\mathbf{t}' + \mathbf{t}''}{2}\right) - \frac{g(\mathbf{t}') + g(\mathbf{t}'')}{2}.$$

The above yields a contradiction to the strict convexity of g on C_g . \square

Using the result from Lemma 3.4 above, we now proceed to prove that $\{\mathbf{t}^{r,i}\}_{r \geq 0}$ converges to \mathbf{t}^* for every $0 \leq i \leq n$, and then use this to establish that the sequence of differences between the successive iterates produced by the cyclic coordinatewise descent algorithm for f_1 is square-summable.

Lemma 3.5 (a) For every $0 \leq i \leq n$,

$$\|\mathbf{t}^{r,i} - \mathbf{t}^*\| \rightarrow 0, \quad (3.16)$$

as $r \rightarrow \infty$.

(b)

$$\sum_{r=0}^{\infty} \|\mathbf{x}^r - \mathbf{x}^{r+1}\|^2 < \infty.$$

Proof (a) Fix i between 0 to n arbitrarily. Since $\{\mathbf{t}^{r,i}\}_{r \geq 0}$ is bounded, it has at least one limit point. Let \mathbf{t}^∞ be an arbitrarily chosen limit point. Hence, there exists a subsequence \mathcal{R} of \mathbb{N} such that $\{\mathbf{t}^{r,i}\}_{r \in \mathcal{R}}$ converges to \mathbf{t}^∞ . Note that $\mathbf{t}^\infty \in C_g$. Hence, g is continuously differentiable in an open set around \mathbf{t}^∞ .

Note that for every $j \neq i$,

$$\|\mathbf{x}^{r,j} - \mathbf{x}^{r,i}\| = \sqrt{\sum_{k=\min(i,j)+1}^{\max(i,j)} |x_k^{r+1} - x_k^r|^2} \leq \|\mathbf{x}^{r+1} - \mathbf{x}^r\|.$$

It follows by Lemma 3.4 that $\|\mathbf{x}^{r,j} - \mathbf{x}^{r,i}\| \rightarrow 0$ as $r \rightarrow \infty$ for every $0 \leq j \leq n$. Hence, we have $\|\mathbf{t}^{r,j} - \mathbf{t}^{r,i}\| \rightarrow 0$ as $r \rightarrow \infty$ for every $0 \leq j \leq n$. It follows that

$$\{\mathbf{t}^{r,j}\}_{r \in \mathcal{R}} \rightarrow \mathbf{t}^\infty \quad (3.17)$$

for every $0 \leq j \leq n$. Let $d^\infty = E^T \nabla g(\mathbf{t}^\infty)$. It follows that

$$\{d(\mathbf{x}^{r,j})\}_{r \in \mathcal{R}} = \{E^T \nabla g(\mathbf{t}^{r,j})\}_{r \in \mathcal{R}} \rightarrow d^\infty \quad (3.18)$$

as for every $0 \leq j \leq n$. By (3.8) and (3.9), it follows that for every $r \in \mathcal{R}$,

$$x_i^{r+1} = x_i^{r,i} = \text{sign}(x_i^{r,i} - d_i(\mathbf{x}^{r,i})) \max(|x_i^{r,i} - d_i(\mathbf{x}^{r,i})| - \lambda, 0), \quad (3.19)$$

for $i \in S$, and

$$x_i^{r+1} = x_i^{r,i} = \max(0, x_i^{r,i} - d_i(\mathbf{x}^{r,i})). \quad (3.20)$$

for $i \in S^c$. By the arguments in the proof of Lemma 3.2, it follows that

$$d_i(\mathbf{x}^{r,i}) + \lambda \text{sign}(x_i^{r,i}) = 0 \quad \text{if } x_i \neq 0, i \in S, \quad (3.21)$$

$$|d_i(\mathbf{x}^{r,i})| \leq \lambda \quad \text{if } x_i = 0, i \in S. \quad (3.22)$$

It follows from (3.18) and (3.19) that $|d_i^\infty| \leq \lambda$ for $i \in S$. Since $x_i^{r,i} \geq 0$ for $i \in S^c$, it follows from (3.18) and (3.20) that $d_i^\infty \geq 0$ for $i \in S^c$. If $i \in S$ and $|d_i^\infty| < \lambda$, then $|d_i(\mathbf{x}^{r,i})| < \lambda$ for large enough r . It follows that

$$x_i^{r+1} = x_i^{r,i} = 0. \quad (3.23)$$

If $i \in S^c$ and $d_i^\infty > 0$, then $d_i(\mathbf{x}^{r,i}) > 0$ for large enough r . It follows that

$$x_i^{r+1} = x_i^{r,i} = 0. \quad (3.24)$$

For each $r \in \mathcal{R}$, consider the linear system

$$E\mathbf{x} = \mathbf{t}^{r+1}, \quad x_j = x_j^{r+1} \quad \forall j \in S \text{ and } j \in S^c \text{ with } d_j^\infty > 0, \quad \mathbf{x} \in \mathcal{X}. \quad (3.25)$$

This is a consistent system of equations since \mathbf{x}^{r+1} is a solution. Fix any $\bar{\mathbf{x}} \in \mathcal{X}$. By Lemma 3.3, for every $r \in \mathcal{R}$, there exists a solution \mathbf{y}^r of this linear system satisfying

$$\|\bar{\mathbf{x}} - \mathbf{y}^r\| \leq \theta \left(\|E\bar{\mathbf{x}} - \mathbf{t}^{r+1}\| + \sum_{j \in S} |\bar{x}_j - x_j^{r+1}| + \sum_{j \in S^c: d_j^\infty > 0} |\bar{x}_j - x_j^{r+1}| \right), \quad (3.26)$$

where θ is a constant depending on E only. Note that by (3.10), $\{f_1(\mathbf{x}^{r,i})\}_{r \in \mathcal{R}, 1 \leq i \leq n}$ is bounded above. By (3.11), we get that $\{\mathbf{t}^{r,i}\}_{r \in \mathcal{R}, 1 \leq i \leq n}$ is bounded. Hence, $\{g(\mathbf{t}^{r,i})\}_{r \in \mathcal{R}, 1 \leq i \leq n}$ is bounded below. It follows by the definition of f that $\{\sum_{j \in S} |x_j^{r,i}|\}_{r \in \mathcal{R}, 1 \leq i \leq n}$ is bounded above. Hence, the right hand side of (3.26) is bounded for all $r \in \mathcal{R}$. It follows by (3.17), (3.23) and (3.24) that $\{\mathbf{y}^r\}_{r \in \mathcal{R}}$ is bounded, and that every limit point of $\{\mathbf{y}^r\}_{r \in \mathcal{R}}$, say \mathbf{y}^∞ , satisfies

$$E\mathbf{y}^\infty = \mathbf{t}^\infty, \quad y_j^\infty = 0 \quad \forall j \in S \text{ with } |d_j^\infty| < \lambda \text{ and } j \in S^c \text{ with } d_j^\infty > 0, \quad \mathbf{y}^\infty \in \mathcal{X}. \quad (3.27)$$

Since $E\mathbf{y}^\infty = \mathbf{t}^\infty$, we obtain $d(\mathbf{y}^\infty) = E^T \nabla g(\mathbf{t}^\infty) = d^\infty$.

Note that if $j \in S$ and $d_j^\infty = \lambda$, it follows from (3.18), (3.21) and (3.22) that $x_j^{r+1} = x_j^{r,j} \leq 0$ for large enough r . Since \mathbf{y}^r satisfies (3.25), and \mathbf{y}^∞ is a limit point, it follows that $y_j^\infty \leq 0$. Hence,

$$y_j^\infty = \text{sign}(y_j^\infty - d_j(\mathbf{y}^\infty)) \max(|y_j^\infty - d_j(\mathbf{y}^\infty)| - \lambda, 0). \quad (3.28)$$

If $j \in S$ and $d_j^\infty = -\lambda$, a similar argument as above implies that y_j^∞ satisfies (3.28). If $j \in S^c$ and $d_j^\infty = d_j(\mathbf{y}^\infty) > 0$, then it follows by (3.24) that $\mathbf{y}^i \text{ nfty } y_j = 0$. Hence,

$$y_j^\infty = \max(y_j^\infty - d_j(\mathbf{y}^\infty), 0). \quad (3.29)$$

If $j \in S^c$ and $d_j^\infty > 0$, then (3.29) holds trivially. It follows from Lemma 3.2, (3.27), (3.28) and (3.29) that $\mathbf{y}^\infty \in \mathcal{X}^*$. It follows by Lemma 3.1 that $\mathbf{t}^\infty = E\mathbf{y}^\infty = \mathbf{t}^*$. Since \mathbf{t}^∞ is an arbitrarily chosen limit point of $\{\mathbf{t}^{r,i}\}_{r \geq 0}$, it follows that for every $0 \leq i \leq n$,

$$\|\mathbf{t}^{r,i} - \mathbf{t}^*\| \rightarrow 0$$

as $r \rightarrow \infty$. This establishes part (a).

(b) By Lemma 3.5, it follows that for r sufficiently large, $E\mathbf{x}^{r,i} \in \mathcal{U}^*$ for every $0 \leq i \leq n$. Consider any such r . For every $1 \leq i \leq n$, a second order Taylor series expansion along the i^{th} coordinate leads to

$$g(E\mathbf{x}^{r,i-1}) - g(E\mathbf{x}^{r,i}) = \nabla g(E\mathbf{x}^{r,i})^T E_{\cdot i} (x_i^{r,i-1} - x_i^{r,i}) + E_{\cdot i}^T \nabla^2 g(E\tilde{\mathbf{x}}^{r,i}) E_{\cdot i} (x_i^{r,i-1} - x_i^{r,i})^2,$$

where $\tilde{\mathbf{x}}^{r,i}$ is a convex combination of $\mathbf{x}^{r,i-1}$ and $\mathbf{x}^{r,i}$. Since \mathcal{U}^* is an open ball containing $\mathbf{t}^{r,i-1} = E\mathbf{x}^{r,i-1}$ and $\mathbf{t}^{r,i} = E\mathbf{x}^{r,i}$, we conclude that $E\tilde{\mathbf{x}}^{r,i}$ is contained in \mathcal{U}^* . Note that $d_i(\mathbf{x}^{r,i}) = \nabla g(E\mathbf{x}^{r,i})^T E_{\cdot i}$. It follows from (3.1) that

$$g(E\mathbf{x}^{r,i-1}) - g(E\mathbf{x}^{r,i}) \geq d_i(\mathbf{x}^{r,i})(x_i^{r,i-1} - x_i^{r,i}) + \sigma \|E_{\cdot i}\|^2 (x_i^{r,i-1} - x_i^{r,i})^2. \quad (3.30)$$

for every $1 \leq i \leq n$.

Fix $i \in S$ arbitrarily. Let $h(x) := |x|$. For any subderivative δ of the function h at $x_i^{r,i}$, we have

$$|x_i^{r,i-1}| - |x_i^{r,i}| \geq \delta(x_i^{r,i-1} - x_i^{r,i}). \quad (3.31)$$

Note that $\mathbf{x}^{r,i-1}$ and $\mathbf{x}^{r,i}$ only differ in the i^{th} coordinate. Using the definition of f along with (3.30) and (3.31), it follows that

$$f(\mathbf{x}^{r,i-1}) - f(\mathbf{x}^{r,i}) \geq (d_i(\mathbf{x}^{r,i}) + \lambda\delta)(x_i^{r,i-1} - x_i^{r,i}) + \sigma\|E_{\cdot i}\|^2(x_i^{r,i-1} - x_i^{r,i})^2.$$

Note that if $x_i^{r,i} \neq 0$, then $\delta = \text{sign}(x_i^{r,i})$. If $x_i^{r,i} = 0$, then any $\delta \in [-1, 1]$ is a valid subderivative choice for h , and $\mathbf{x}^{r,i}$ is obtained from $\mathbf{x}^{r,i-1}$ by minimizing f along the i^{th} coordinate. Using these observations, we conclude that it is always possible to choose δ such that $d_i(\mathbf{x}^{r,i}) + \lambda\delta = 0$. Hence, for every $i \in S$

$$f(\mathbf{x}^{r,i-1}) - f(\mathbf{x}^{r,i}) \geq \sigma \left(\min_{1 \leq j \leq n} \|E_{\cdot j}\|^2 \right) (x_i^{r,i-1} - x_i^{r,i})^2. \quad (3.32)$$

Fix $i \in S^c$ arbitrarily. By (3.9), $d_i(\mathbf{x}^{r,i}) \geq 0$. Suppose $d_i(\mathbf{x}^{r,i}) = 0$. Since $\mathbf{x}^{r,i-1}$ and $\mathbf{x}^{r,i}$ only differ in the i^{th} coordinate, it follows by (3.30), the definition of f that (3.32) holds in this case. Suppose $d_i(\mathbf{x}^{r,i}) > 0$. By (3.9), $x_i^{r,i} = 0$. Since $x_i^{r,i-1} \geq 0$, it follows that $d_i(\mathbf{x}^{r,i})(x_i^{r,i-1} - x_i^{r,i}) \geq 0$. Hence, (3.32) holds in this case.

Adding (3.32) over $i = 1, 2, \dots, n$, we obtain

$$f(\mathbf{x}^r) - f(\mathbf{x}^{r+1}) \geq \sigma \left(\min_{1 \leq j \leq n} \|E_{\cdot j}\|^2 \right) \sum_{i=1}^n (x_i^{r,i-1} - x_i^{r,i})^2 = \sigma \left(\min_{1 \leq j \leq n} \|E_{\cdot j}\|^2 \right) \|\mathbf{x}^r - \mathbf{x}^{r+1}\|^2.$$

The result follows by noting that $f(\mathbf{x}^r) \downarrow f^\infty > -\infty$ as $r \rightarrow \infty$ and that $\min_{1 \leq j \leq n} \|E_{\cdot j}\|^2 > 0$ as E has no zero column. \square

Although the square-summability, established above, is an important step towards proving convergence, further arguments are needed to establish convergence of the sequence of iterates generated by Algorithm 1. It follows by Lemma 3.5 and the continuity of ∇g at \mathbf{t}^* that

$$\mathbf{d}(\mathbf{x}^{r,i}) \rightarrow \mathbf{d}^* \quad (3.33)$$

as $r \rightarrow \infty$ for every $1 \leq i \leq n$. The next lemma establishes that for each i , x_i^r has the same sign for sufficiently large r .

Lemma 3.6 (a) For all r sufficiently large, $x_i^r = 0$ for all $i \in S$ with $|d_i^*| < \lambda$ and for all $i \in S^c$ with $d_i^* > 0$.

(b) For all r sufficiently large, $x_i^r \leq 0$ for all $i \in S$ with $d_i^* = \lambda$, and $x_i^r \geq 0$ for all $i \in S$ with $d_i^* = -\lambda$.

Proof The proof of part (a) follows by using exactly the same arguments as those leading to (3.23) and (3.24). We now prove part (b). Let $i \in S$ with $d_i^* = \lambda$. Note that by (3.8), $x_i^r = \text{sign}(x_i^r - d_i(\mathbf{x}^{r-1,i})) \max(|x_i^r - d_i(\mathbf{x}^{r-1,i})| - \lambda, 0)$. By exactly the same arguments as

in the proof of Lemma 3.2, it follows that $d_i(\mathbf{x}^{r-1,i}) + \lambda \text{sign}(x_i^r) = 0$ if $x_i^r \neq 0$. By (3.33), it follows that $d_i(\mathbf{x}^{r-1,i}) \rightarrow \lambda$ as $r \rightarrow \infty$. Hence, for sufficiently large r , $x_i^r \neq 0$ implies that $x_i^r > 0$. The other case follows similarly. \square

For every $\mathbf{x} \in \mathcal{X}$, define the function ϕ as follows:

$$\phi(\mathbf{x}) = \min_{\mathbf{x}^* \in \mathcal{X}^*} \|\mathbf{x} - \mathbf{x}^*\|.$$

Hence, $\phi(\mathbf{x})$ is the distance of \mathbf{x} from the closed convex set \mathcal{X}^* . The goal of the next lemma is to establish that the sequence of iterates $\{\mathbf{x}^r\}_{r \geq 0}$ approaches the boundary of \mathcal{X}^* . Although, this lemma is a useful component of the convergence proof of the CCM algorithm for f_1 , it clearly is not sufficient to establish convergence. For example, consider a sequence which alternatively takes two distinct values at the boundary of a set. It easily follows that the distance of the sequence from the boundary of that set is always zero, but the sequence still does not converge.

Lemma 3.7 (a) *Let $\lambda > 0$. Then*

$$|\text{sign}(a) \max(|a| - \lambda, 0) - \text{sign}(b) \max(|b| - \lambda, 0)| \leq |a - b|, \quad (3.34)$$

for all $a, b \in \mathbb{R}$.

(b) *If $i \in S$, then*

$$x_i^r - \text{sign}(x_i^r - d_i(\mathbf{x}^r)) \max(|x_i^r - d_i(\mathbf{x}^r)| - \lambda, 0) \rightarrow 0$$

as $r \rightarrow \infty$. If $i \in S^c$, then

$$x_i^r - \max(x_i^r - d_i(\mathbf{x}^r), 0) \rightarrow 0$$

as $r \rightarrow \infty$.

(c)

$$\phi(\mathbf{x}^r) \rightarrow 0 \text{ as } r \rightarrow \infty.$$

Proof (a) We consider various cases:

Suppose $|a| \leq \lambda$ and $|b| \leq \lambda$. Then the $\max(|a| - \lambda, 0) = \max(|b| - \lambda, 0) = 0$. Hence, (3.34) holds. Suppose $|a| \leq \lambda$ and $|b| > \lambda$. Then

$$\begin{aligned} |\text{sign}(a) \max(|a| - \lambda, 0) - \text{sign}(b) \max(|b| - \lambda, 0)| &= |b| - \lambda \\ &\leq |b| - |a| \\ &\leq |b - a|. \end{aligned}$$

The last step follows by the triangle inequality. The case $|a| > \lambda$ and $|b| \leq \lambda$ can be analyzed similarly. Next, suppose that $|a| > \lambda$, $|b| > \lambda$ and $\text{sign}(a) = \text{sign}(b)$. Then

$$\begin{aligned} |\text{sign}(a) \max(|a| - \lambda, 0) - \text{sign}(b) \max(|b| - \lambda, 0)| &= |a - b + \lambda \text{sign}(b) - \lambda \text{sign}(a)| \\ &= |a - b|. \end{aligned}$$

Finally, we consider the case when $|a| > \lambda$, $|b| > \lambda$ and $\text{sign}(a) \neq \text{sign}(b)$. Without loss of generality, let $a < 0$ and $b > 0$. Then

$$\begin{aligned} |\text{sign}(a) \max(|a| - \lambda, 0) - \text{sign}(b) \max(|b| - \lambda, 0)| &= |a - b + 2\lambda| \\ &= b + |a| - 2\lambda \\ &< b + |a| \\ &= |a - b|. \end{aligned}$$

(b) Note that by (3.8), $x_i^r = \text{sign}(x_i^r - d_i(\mathbf{x}^{r-1,i})) \max(|x_i^r - d_i(\mathbf{x}^{r-1,i})| - \lambda, 0)$. It follows by Lemma 3.6, part (a) of this lemma, and (3.33) that

$$|x_i^r - \text{sign}(x_i^r - d_i(\mathbf{x}^r)) \max(|x_i^r - d_i(\mathbf{x}^r)| - \lambda, 0)| \leq |d_i(\mathbf{x}^{r-1,i}) - d_i(\mathbf{x}^r)| \rightarrow 0$$

as $r \rightarrow \infty$. Similarly, by (3.9), $x_i^r = \max(x_i^r - d_i(\mathbf{x}^{r-1,i}), 0)$. Note that for any $a, b \in \mathbb{R}$, $|\max(a, 0) - \max(b, 0)| < |a - b|$. It follows by (3.33) that

$$|x_i^r - \max(x_i^r - d_i(\mathbf{x}^r), 0)| \leq |d_i(\mathbf{x}^{r-1,i}) - d_i(\mathbf{x}^r)| \rightarrow 0.$$

as $r \rightarrow \infty$.

(c) By Lemma 3.1, (3.3), (3.4) and (3.5), it follows that \mathcal{X}^* is the solution set of the linear system of equations given by

$$\begin{aligned} E\mathbf{y} &= \mathbf{t}^*, \mathbf{y} \in \mathcal{X}, \\ y_i &= 0 \text{ if } i \in S, |d_i^*| < \lambda, \\ y_i &\leq 0 \text{ if } i \in S, d_i^* = \lambda, \\ y_i &\geq 0 \text{ if } i \in S, d_i^* = -\lambda, \\ y_i &= 0 \text{ if } i \in S^c, d_i^* > 0. \end{aligned}$$

Since \mathcal{X}^* is non-empty, by Lemma 3.3 and Lemma 3.6, for sufficiently large r , there exists $\mathbf{y}^r \in \mathcal{X}^*$ such that

$$\begin{aligned} \|\mathbf{x}^r - \mathbf{y}^r\| &\leq \theta \left(\|E\mathbf{x}^r - \mathbf{t}^*\| + \sum_{i \in S, d_i^* = \lambda} (x_i^r)^+ + \sum_{i \in S, d_i^* = -\lambda} (-x_i^r)^+ + \sum_{i \in S, |d_i^*| < \lambda} |x_i^r| \right) + \\ &\quad \theta \sum_{i \in S^c, d_i^* > 0} |x_i^r| \\ &= \theta \|E\mathbf{x}^r - \mathbf{t}^*\| \end{aligned} \tag{3.35}$$

where θ is a constant only depending on E . The result follows by the definition of ϕ and Lemma 3.5. \square

Let

$$\begin{aligned} I_1^* &:= \{i \in S : d_i^* = \lambda\}, \\ I_2^* &:= \{i \in S : d_i^* = -\lambda\}, \\ I_3^* &:= \{i \in S : |d_i^*| < \lambda\}, \\ I_4^* &:= \{i \in S^c : d_i^* = 0\}, \\ I_5^* &:= \{i \in S^c : d_i^* > 0\}. \end{aligned}$$

For $\mathbf{x} \in \mathbb{R}^n$ and $M \subseteq \{1, 2, \dots, n\}$, let $\mathbf{x}_M := (x_i)_{i \in M}$. By Lemma 3.6 and (3.9) that there exists an $r_0 > 0$ such that

$$\mathbf{x}_{I_3^* \cup I_5^*}^r = \mathbf{0}, (\mathbf{x}_{I_1^*}^r)^+ = \mathbf{0}, (-\mathbf{x}_{I_2^* \cup I_4^*}^r)^+ = \mathbf{0} \quad (3.36)$$

for every $r \geq r_0$. The following lemma provides a crucial identity which will play an important role in the last leg of the convergence proof.

Lemma 3.8 *There exists $\omega > 0$ such that*

$$\|E\mathbf{x}^r - \mathbf{t}^*\| \leq \omega \|\mathbf{x}^r - \mathbf{x}^{r+1}\|$$

for every $r \geq r_0$.

Proof Consider arbitrary (possibly empty) subsets I_1, I_2, I_3 of S and I_4, I_5 of S^c , and let \mathcal{R} denote the set of indices $r \geq r_0$ for which

$$d_i(\mathbf{x}^{r,i}) = \lambda \quad \forall i \in I_1, \quad (3.37)$$

$$d_i(\mathbf{x}^{r,i}) = -\lambda \quad \forall i \in I_2, \quad (3.38)$$

$$|d_i(\mathbf{x}^{r,i})| < \lambda \quad \forall i \in I_3, \quad (3.39)$$

$$d_i(\mathbf{x}^{r,i}) = 0 \quad \forall i \in I_4, \quad (3.40)$$

$$d_i(\mathbf{x}^{r,i}) > 0 \quad \forall i \in I_5. \quad (3.41)$$

Note that $|d_i(\mathbf{x}^{r,i})| \leq \lambda$ for $i \in S$, and $d_i(\mathbf{x}^{r,i}) \geq 0$ for $i \in S^c$. Hence, $I_1 \cup I_2 \cup I_3 = S$ and $I_4 \cup I_5 = S^c$. Suppose we are able to show that there exists a constant $\omega_{I_1, I_2, I_3, I_4, I_5} > 0$ such that

$$\|E\mathbf{x}^r - \mathbf{t}^*\| \leq \omega_{I_1, I_2, I_3, I_4, I_5} \|\mathbf{x}^r - \mathbf{x}^{r+1}\|, \quad (3.42)$$

for every $r \in \mathcal{R}$. Since every $r \geq r_0$ belongs to \mathcal{R} corresponding to some choice of $\{I_j\}_{1 \leq j \leq 5}$, and the number of distinct choices of $\{I_j\}_{1 \leq j \leq 5}$ is finite, it would immediately imply that the lemma holds with

$$\omega = \max_{I_1, I_2, I_3, I_4, I_5} \omega_{I_1, I_2, I_3, I_4, I_5}.$$

Hence, we now establish (3.42). Note that if $\mathbf{x}^{r+1} = \mathbf{x}^r$, then by Lemma 3.1, Lemma 3.2, (3.8) and (3.9), it follows that $\mathbf{x}^r \in \mathcal{X}^*$ and $E\mathbf{x}^r = \mathbf{t}^*$. Hence, if \mathcal{R} is empty or finite, then the result holds trivially. Hence, we assume that \mathcal{R} is infinite. It follows by (3.8), (3.9) and (3.37)-(3.41) that

$$\mathbf{x}_{I_3 \cup I_5}^{r+1} = \mathbf{0}, (\mathbf{x}_{I_1}^{r+1})^+ = \mathbf{0}, (-\mathbf{x}_{I_2 \cup I_4}^{r+1})^+ = \mathbf{0}. \quad (3.43)$$

Consider the linear system

$$\mathbf{y}_{I_3 \cup I_5} = \mathbf{0}, \mathbf{y} \in \mathcal{X}^*. \quad (3.44)$$

By an argument very similar to the one following [20, eq. (B.7)], it follows that the above linear system is consistent (essentially by noting that $\{\mathbf{y} \in \mathcal{X} : \mathbf{y}_{I_3 \cup I_5} = \mathbf{0}\}$ and \mathcal{X}^* are polyhedral sets, and proving that they get arbitrarily close to each other). It follows by

Lemma 3.1, (3.3), (3.4) and (3.5) that the solution set of the linear system in (3.44) is identical to the solution set of the following linear system.

$$\begin{aligned} \mathbf{y}_{I_3 \cup I_5} &= \mathbf{0}, \\ E\mathbf{y} &= \mathbf{t}^*, \mathbf{y} \in \mathcal{X}, \\ \mathbf{y}_{I_3^* \cup I_5^*} &= \mathbf{0}, \\ (\mathbf{y}_{I_1^*})^+ &= \mathbf{0}, \\ (-\mathbf{y}_{I_2^* \cup I_4^*})^+ &= \mathbf{0}. \end{aligned}$$

It follows by Lemma 3.3 that, for every $r \in \mathcal{R}$, there exists a solution \mathbf{y}^r to the above linear system satisfying

$$\|\mathbf{x}^r - \mathbf{y}^r\| \leq \kappa_1(\|E\mathbf{x}^r - \mathbf{t}^*\| + \|\mathbf{x}_{I_3 \cup I_3^* \cup I_5 \cup I_5^*}^r\| + \|(\mathbf{x}_{I_1^*}^r)^+\| + \|(-\mathbf{x}_{I_2^* \cup I_4^*}^r)^+\|), \quad (3.45)$$

where κ_1 depends only on E . It follows by (3.36) and (3.43) that

$$\begin{aligned} \|\mathbf{x}^r - \mathbf{y}^r\| &\leq \kappa_1(\|E\mathbf{x}^r - \mathbf{t}^*\| + \|\mathbf{x}_{I_3 \cup I_3^* \cup I_5 \cup I_5^*}^r - \mathbf{x}_{I_3 \cup I_3^* \cup I_5 \cup I_5^*}^{r+1}\|) \\ &\leq \kappa_1(\|E\mathbf{x}^r - \mathbf{t}^*\| + \|\mathbf{x}^r - \mathbf{x}^{r+1}\|). \end{aligned} \quad (3.46)$$

For any $m \times n$ matrix A and $M \subseteq \{1, 2, \dots, n\}$, let $A_M := ((A_{ij}))_{i \in \{1 \leq i \leq m, j \in M\}}$. Let $I = I_1 \cup I_2 \cup I_4$. Note that by (3.43) $\|E_{I^c}(\mathbf{x}_{I^c}^r - \mathbf{y}_{I^c}^r)\| = \|E_{I^c}(\mathbf{x}_{I^c}^r - \mathbf{x}_{I^c}^{r+1})\| \leq \|E\|\|\mathbf{x}^r - \mathbf{x}^{r+1}\|$. It follows by Lemma 3.1 and (3.46) that

$$\begin{aligned} \|\mathbf{x}_{I^c}^r - \mathbf{y}_{I^c}^r\| &\leq \|\mathbf{x}^r - \mathbf{y}^r\| \\ &\leq \kappa_1(\|E(\mathbf{x}^r - \mathbf{y}^r)\| + \|\mathbf{x}^r - \mathbf{x}^{r+1}\|) \\ &\leq \kappa_1((1 + \|E\|)\|\mathbf{x}^r - \mathbf{x}^{r+1}\| + \|E_I(\mathbf{x}_I^r - \mathbf{y}_I^r)\|) \\ &\leq \kappa_1(1 + \|E\|)(\|\mathbf{x}^r - \mathbf{x}^{r+1}\| + \|E_I(\mathbf{x}_I^r - \mathbf{y}_I^r)\|) \end{aligned} \quad (3.47)$$

Let $\mathbf{c} \in \mathbb{R}^n$ be such that $\mathbf{c}_{I_1} = \lambda$, $\mathbf{c}_{I_2} = -\lambda$, and all the other entries of \mathbf{c} are equal to zero. It follows by Lemma 3.1, (3.37), (3.38) and (3.40) that

$$\|\mathbf{d}_I(\mathbf{x}^r) - \mathbf{c}_I\| = \|\mathbf{d}_I(\mathbf{x}^r) - \mathbf{d}_I^*\| = \|\mathbf{d}_I(\mathbf{x}^r) - \mathbf{d}_I(\mathbf{y}^r)\| = \|(E_I)^T \nabla g(E\mathbf{x}^r) - (E_I)^T \nabla g(E\mathbf{y}^r)\|,$$

and

$$|d_i(\mathbf{x}^r) - c_i| = |d_i(\mathbf{x}^r) - d_i(\mathbf{x}^{r,i})| = |E_{i,i}^T \nabla g(E\mathbf{x}^r) - E_{i,i}^T \nabla g(E\mathbf{x}^{r,i})|.$$

for every $i \in I$. The result now follows by exactly the same arguments as in [20] (from [20, eq. (B.9)] to the end of the proof of [20, Lemma B.3], using $\|\mathbf{d}_I(\mathbf{x}^r) - \mathbf{c}_I\|$ in place of $\|\mathbf{d}_I(\mathbf{x}^r)\|$, and replacing E by E^T throughout). \square

We now invoke two matrix-theoretic results from Luo and Tseng [20]. Let $M = E^T \nabla^2 g(\mathbf{t}^*) E$. By (A4) and the assumption that E has no zero column, it follows that $m_{ii} > 0$ for every $1 \leq i \leq n$. For any $J, \tilde{J} \subseteq \{1, 2, \dots, n\}$, let $M_{J\tilde{J}} := (M_{ij})_{i \in J, j \in \tilde{J}}$, and $|J|$ denote the cardinality of J . The following lemma is provided in Luo and Tseng [20], and exploits the fact that M is symmetric positive semi-definite.

Lemma 3.9 (Luo and Tseng [20]) *Let $J \subseteq \{1, 2, \dots, n\}$. Then $\text{Span}(M_{JJ^c}) \subseteq \text{Span}(M_{JJ})$.*

Let B denote the lower triangular portion of M , and $C = M - B$ denote the strictly upper triangular portion of M (hence the diagonal entries of C are zero). We use the following lemma from [20].

Lemma 3.10 (Luo and Tseng [20]) (a) *For any nonempty $J \subseteq \{1, 2, \dots, n\}$, there exist $\rho_J \in (0, 1)$ and $\tau_J > 0$ such that*

$$\|(I - M_{JJ}(B_{JJ})^{-1})^k \mathbf{z}\| \leq \tau_J(\rho_J)^k \|\mathbf{z}\|, \quad \forall k \geq 1, \quad \forall \mathbf{z} \in \text{Span}(M_{JJ}).$$

(b) *There exists a $\Delta \geq 1$ such that, for any nonempty $J \subseteq \{1, 2, \dots, n\}$,*

$$\|(I - M_{JJ}(B_{JJ})^{-1})^k \mathbf{z}\| \leq \Delta \|\mathbf{z}\|, \quad \forall k \geq 1, \quad \forall \mathbf{z} \in \mathbb{R}^{|J|}.$$

Let $I^* = I_1^* \cup I_2^* \cup I_4^*$, and

$$\beta = \max_{J \subseteq I^*} \sqrt{|J^c|} \left\{ \left(\frac{\tau_J \|(B_{JJ}^{-1})\| \|M_{JJ}\|}{1 - \rho_J} + \Delta + 1 \right) \|(B_{JJ})^{-1} B_{JJ^c}\| + \frac{\tau_J \|(B_{JJ})^{-1}\| \|M_{JJ}\|}{1 - \rho_J} \right\}.$$

Recall that by Lemma 3.6 and (3.9), there exists an $r_0 > 0$ such that

$$\mathbf{x}_{I_3^* \cup I_5^*}^r = \mathbf{0}, \quad (\mathbf{x}_{I_1^*}^r)^+ = \mathbf{0}, \quad (-\mathbf{x}_{I_2^* \cup I_4^*}^r)^+ = \mathbf{0}$$

for every $r \geq r_0$. For $\mathbf{x} \in \mathbb{R}^n$, let $\|\mathbf{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|$. The next lemma is analogous to Lemma 9 of [20], and shows that the coordinates of \mathbf{x}^r that stay away from zero, are influenced by the coordinates which eventually become zero only through the distance of these coordinates from zero.

Lemma 3.11 *Consider any $J \subseteq I^*$. If for some two integers $s \geq t \geq r_0$ we have $x_i^r \neq 0$ for every $t + 1 \leq r \leq s$ and $i \in J$, then, for any $\mathbf{x}^* \in \mathcal{X}^*$, there holds*

$$\|\mathbf{x}_J^s - \mathbf{x}_J^*\| \leq \Delta \|\mathbf{x}_J^t - \mathbf{x}_J^*\| + \beta \max_{t \leq r \leq s} \|\mathbf{x}_{J^c}^r - \mathbf{x}_{J^c}^*\|_\infty + \mu \sum_{r=t}^{s-1} \|\mathbf{x}^r - \mathbf{x}^{r+1}\|^2,$$

where μ is some positive constant which is independent of s and t .

The proof of the lemma above is provided in the appendix. Let $\sigma_0 := 1$ and

$$\sigma_k = \Delta + 3 + \beta + (\beta + 1)\sigma_{k-1} + \mu, \quad k = 1, 2, \dots, n.$$

It follows from the above definition that $\sigma_k \geq 1$ for every $1 \leq k \leq n$, and is monotonically increasing with k .

Fix $\delta > 0$ arbitrarily. By Lemma 3.4, Lemma 3.5 and Lemma 3.7, there exists $r_1 > 0$ such that

$$\phi(\mathbf{x}^r) \leq \delta, \tag{3.48}$$

$$\|\mathbf{x}^{r+1} - \mathbf{x}^r\| \leq \delta, \tag{3.49}$$

$$\sum_{k=r}^{\infty} \|\mathbf{x}^k - \mathbf{x}^{k+1}\|^2 \leq \delta, \tag{3.50}$$

for every $r \geq r_1$.

The next three lemmas are analogous to [19, Lemma 10], [19, Lemma 11], and [19, Lemma 9] respectively. The crucial difference is that we consider absolute values of appropriate vector entries (as opposed to the lemmas in [19], which use the vector entries themselves). Recall that $I^* = I_1^* \cup I_2^* \cup I_4^*$. The proofs of all three lemmas are provided in the appendix.

Lemma 3.12 *Fix $k \in \{1, 2, \dots, n\}$ arbitrarily. If for some nonempty $J \subset I^*$, and some intergers $t' > t \geq \max(r_0, r_1)$, we have*

$$|x_i^t| > \sigma_k \delta, \forall i \in J, \quad (3.51)$$

$$|x_i^r| \leq \sigma_{k-1} \delta, \forall i \notin J, \forall r = t, t+1, \dots, t'-1, \quad (3.52)$$

then the following hold:

(a) $|x_i^{t'}| > \sigma_{k-1} \delta$ for every $i \in J$.

(b) *There exists an $\mathbf{x}^* \in \mathcal{X}^*$ such that*

$$\|\mathbf{x}^r - \mathbf{x}^*\|_\infty \leq \sigma_k \delta, \forall r = t, t+1, \dots, t'-1.$$

The next lemma extends the previous lemma by removing the assumption that the coordinates that start near zero remain near zero.

Lemma 3.13 *Fix $k \in \{1, 2, \dots, n\}$ arbitrarily. If for some $J \subseteq I^*$ with $|J| \geq |I^*| - k + 1$ and some interger $t > \max(r_0, r_1)$ we have*

$$|x_i^t| > \sigma_k \delta, \forall i \in J, \quad (3.53)$$

$$|x_i^t| \leq \sigma_{k-1} \delta, \forall i \notin J, \quad (3.54)$$

then there exists an $\mathbf{x}^ \in \mathcal{X}^*$ and a $\bar{t} \geq t$ satisfying*

$$\|\mathbf{x}^r - \mathbf{x}^*\|_\infty \leq \sigma_k \delta, \quad (3.55)$$

for every $r \geq \bar{t}$.

We use Lemma 3.13 to establish the final lemma in our analysis.

Lemma 3.14 *For any $\delta > 0$, there exists an $\mathbf{x}^* \in \mathcal{X}^*$ and $\hat{r} > 0$ such that*

$$\|\mathbf{x}^r - \mathbf{x}^*\|_\infty \leq \sigma_n \delta + \delta, \quad (3.56)$$

for every $r \geq \hat{r}$.

Using Lemma 3.14, we are now able to complete the proof of our meta-theorem, Theorem 2.1.

Proof of Theorem 2.1 Fix $\epsilon > 0$ arbitrarily. By Lemma 3.14, there exists $\mathbf{x}^* \in \mathcal{X}^*$ and $\hat{r} > 0$ such that

$$\|\mathbf{x}^r - \mathbf{x}^*\|_\infty < \frac{\epsilon}{2},$$

for every $r \geq \hat{r}$. Hence, for every $r_1, r_2 > \hat{r}$, we obtain by the triangle inequality that

$$\begin{aligned} \|\mathbf{x}^{r_1} - \mathbf{x}^{r_2}\|_\infty &\leq \|\mathbf{x}^{r_1} - \mathbf{x}^*\|_\infty + \|\mathbf{x}^{r_2} - \mathbf{x}^*\|_\infty \\ &< \epsilon. \end{aligned}$$

It follows that the sequence of iterates $\{\mathbf{x}^r\}_{r \geq 0}$ form a Cauchy sequence. By Lemma 3.7, we conclude that $\{\mathbf{x}^r\}_{r \geq 0}$ converges to an element of \mathcal{X}^* . \square

Remark 1 *Note that Theorem 2.1 holds for any $m \times n$ matrix E with non-zero columns, and any subset S of $\{1, 2, \dots, n\}$. It follows that Theorem 2.1 holds for an arbitrary permutation of the order in which the n coordinates are updated in the cyclic coordinatewise descent algorithm.*

4 Convergence analysis of cyclic coordinatewise minimization for f_2

In this section, we consider the convergence behavior of the cyclic coordinatewise minimization algorithm applied to the function f_2 (Algorithm 2). It follows by assumption (A5)* and the convexity of f_2 that the set of optimal solutions of the minimization problem in (1.2), denoted by \mathcal{X}_ℓ^* , is non-empty. Since the negative logarithm function is convex, and q is strictly convex, it follows by arguments very similar to those in [20, Page 5] that \mathcal{X}_ℓ^* is a convex set and that there exists $\mathbf{t}^* \in \mathbb{R}^m$ such that $E\mathbf{x}^* = \mathbf{t}^*$, $\forall \mathbf{x}^* \in \mathcal{X}_\ell^*$. Let $\mathbf{d}(\mathbf{x}) = \nabla\{q(E\mathbf{x})\} = 2E^T E\mathbf{x}$. We denote the i^{th} entry of $\mathbf{d}(\mathbf{x})$ by $d_i(\mathbf{x})$. Let $\mathbf{d}^* := E^T E\mathbf{x}^* = E\mathbf{t}^*$. It follows that

$$\mathbf{d}(\mathbf{x}^*) = \mathbf{d}^* \quad \forall \mathbf{x}^* \in \mathcal{X}_\ell^*. \quad (4.1)$$

We now state two lemmas which will be important in understanding the coordinatewise minimization for the function f_2 .

Lemma 4.1 (a) *Let $h(u) = au^2 + bu + c - \log u$ for $u > 0$. If $a > 0$, then $h(u)$ is uniquely minimized at $u^* = \frac{-b + \sqrt{b^2 + 4a}}{2a}$.*

(b) *Let $h(u) = au^2 + bu + c + \lambda|u|$ for $u \in \mathbb{R}$. If $a, \lambda > 0$, then $h(u)$ is uniquely minimized at $u^* = S_\lambda(-b)/2a$, where S_λ is the soft-thresholding operator defined by $S_\lambda(x) = \text{sign}(x)(|x| - \lambda)_+$.*

Proof (a) Note that

$$\frac{d}{du}h(u) = 0 \Leftrightarrow 2au^2 + bu - 1 = 0.$$

The result follows by noting that u^* the only non-negative solution of the above equation, and that h is a strictly convex function.

(b) The KKT conditions for the minimizing the strictly convex function h are satisfied if and only if $u = 0$ if $|b| \leq \lambda$, and $2au + b + \lambda \text{sign}(u) = 0$ if $|b| > \lambda$, which in turn is satisfied if and only if $u = u^*$. \square

It is clear from Lemma 4.1 that the coordinatewise minimizers for f_2 (see (2.3)) are uniquely defined and can be obtained in closed form. Note that the function $f_2(\mathbf{x})$ takes the value

infinity if $x_i = 0$ for i belonging to a non-trivial subset of S^c . Hence, the KKT conditions for the convex minimization problem in (1.2) imply that $\mathbf{x} \in \mathcal{X}_\ell^*$ if and only if

$$d_i(\mathbf{x}) = \frac{1}{x_i} \text{ for } i \in S^c, \quad (4.2)$$

$$d_i(\mathbf{x}) + \lambda \text{sign}(x_i) = 0 \text{ if } x_i \neq 0, i \in S, \quad (4.3)$$

$$|d_i(\mathbf{x})| \leq \lambda \text{ if } x_i = 0, i \in S. \quad (4.4)$$

The arguments in the proof of Lemma 3.2 can be used to provide the following alternative characterization of the elements of \mathcal{X}_ℓ^* .

Lemma 4.2 $\mathbf{x} \in \mathcal{X}_\ell^*$ if and only if

$$d_i(\mathbf{x}) = \frac{1}{x_i} \text{ for } i \in S^c, \quad (4.5)$$

$$x_i = \text{sign}(x_i - d_i(\mathbf{x})) \max(|x_i - d_i(\mathbf{x})| - \lambda, 0) \text{ for } i \in S. \quad (4.6)$$

Recall that $\{\mathbf{z}^r\}_{r \geq 0}$ is the sequence of iterates generated by Algorithm 2, and $\mathbf{x}^{r,i}$ is the appropriate coordinatewise minimizer defined in (2.3). It follows from arguments similar to those in the proof of Lemma 3.2 that for $i \in S$,

$$z_i^{r,i} = \text{sign}(z_i^{r,i} - d_i(\mathbf{z}^{r,i})) \max(|z_i^{r,i} - d_i(\mathbf{z}^{r,i})| - \lambda, 0), \quad (4.7)$$

and for $i \in S^c$

$$z_i^{r,i} = \frac{1}{d_i(\mathbf{z}^{r,i})}. \quad (4.8)$$

As in Section 3, we will establish a series of lemmas, which will ultimately lead us to the proof of Theorem 2.2. Let

$$\mathbf{t}^{r,i} = E\mathbf{z}^{r,i}$$

for all r and all $0 \leq i \leq n$. By (2.3), it follows that

$$f_2(\mathbf{z}^{r,i}) \leq f_2(\mathbf{z}^{r,i-1}) \quad (4.9)$$

for every r and $1 \leq i \leq n$. It follows by assumption (A5)* that

$$\{\mathbf{t}^{r,i}\}_{r \geq 0, 1 \leq i \leq n} \text{ is bounded.} \quad (4.10)$$

By (4.9), the sequence $\{f_2(\mathbf{z}^{r,i})\}_{r \geq 0}$ decreases to the same quantity, say f^∞ for every $0 \leq i \leq n$. Since \mathcal{X}_ℓ^* is non-empty, it follows that $f^\infty > -\infty$. The next lemma shows that the sum of norm-square of the difference between successive iterates in $\{\mathbf{z}^r\}_{r \geq 0}$ is finite. Note that in Section 3, we first needed to show that $\|\mathbf{z}^r - \mathbf{z}^{r+1}\|$ converges to zero (Lemma 3.4) to prove a similar result (Lemma 3.5). However, since we have to deal with the quadratic function $q(E\mathbf{x})$ as opposed to a general $g(E\mathbf{x})$ in this section, a direct argument is available.

Lemma 4.3

$$\sum_{r=0}^{\infty} \|\mathbf{z}^r - \mathbf{z}^{r+1}\|^2 < \infty.$$

Proof For every $1 \leq i \leq n$, a second order Taylor series expansion along the i^{th} coordinate leads to the following.

$$g(E\mathbf{z}^{r,i-1}) - g(E\mathbf{z}^{r,i}) = d_i(\mathbf{z}^{r,i})(z_i^{r,i-1} - z_i^{r,i}) + 2\|E_{\cdot i}\|^2(z_i^{r,i-1} - z_i^{r,i})^2. \quad (4.11)$$

Fix $i \in S$ arbitrarily. Using exactly the same argument as in the proof of Lemma 3.5 (b) for this case, we get that

$$f(\mathbf{z}^{r,i-1}) - f(\mathbf{z}^{r,i}) \geq 2 \left(\min_{1 \leq j \leq n} \|E_{\cdot j}\|^2 \right) (z_i^{r,i-1} - z_i^{r,i})^2. \quad (4.12)$$

Fix $i \in S^c$ arbitrarily. By strict convexity of the negative logarithm function on \mathbb{R}_+ , it follows that

$$(-\log z_i^{r,i-1}) - (-\log z_i^{r,i}) \geq \left(-\frac{1}{z_i^{r,i}} \right) (z_i^{r,i-1} - z_i^{r,i}). \quad (4.13)$$

It follows by (4.8), (4.11) and (4.13) that (4.12) is satisfied for every $i \in S^c$. Adding (4.8) over $i = 1, 2, \dots, n$, we obtain

$$f(\mathbf{z}^r) - f(\mathbf{z}^{r+1}) \geq 2 \left(\min_{1 \leq j \leq n} \|E_{\cdot j}\|^2 \right) \sum_{i=1}^n (z_i^{r,i-1} - z_i^{r,i})^2 = 2 \left(\min_{1 \leq j \leq n} \|E_{\cdot j}\|^2 \right) \|\mathbf{z}^r - \mathbf{z}^{r+1}\|^2.$$

The result follows by noting that $f(\mathbf{z}^r) \downarrow f^\infty > -\infty$ as $r \rightarrow \infty$ and that $\min_{1 \leq j \leq n} \|E_{\cdot j}\|^2 > 0$ as E has no zero column. \square

By Lemma 4.3, it follows that $\|\mathbf{z}^r - \mathbf{z}^{r+1}\| \rightarrow 0$ as $r \rightarrow \infty$. We now establish a parallel version of Lemma 3.5 for the problem at hand.

Lemma 4.4 *For every $0 \leq i \leq n$,*

$$\|\mathbf{t}^{r,i} - \mathbf{t}^*\| \rightarrow 0, \quad (4.14)$$

as $r \rightarrow \infty$.

Proof By exactly the same set of arguments as in the proof of Lemma 3.5, there exists $\mathbf{t}^\infty \in \mathbb{R}^m$, and a subsequence \mathcal{R} of \mathbb{N} such that

$$\{\mathbf{t}^{r,j}\}_{r \in \mathcal{R}} \rightarrow \mathbf{t}^\infty \quad (4.15)$$

for every $0 \leq j \leq n$. Let $d^\infty = 2E^T \mathbf{t}^\infty$. It follows that

$$\{d(\mathbf{z}^{r,j})\}_{r \in \mathcal{R}} = \{2E^T \mathbf{t}^{r,j}\}_{r \in \mathcal{R}} \rightarrow d^\infty \quad (4.16)$$

as for every $0 \leq j \leq n$. Suppose $i \in S$. By repeating exactly the same arguments in the proof of Lemma 3.5 in this case, we get the following.

- If $|d_i^\infty| < \lambda$, then

$$z_i^{r+1} = z_i^{r,i} = 0 \quad (4.17)$$

for large enough r .

- If $d_i^\infty = \lambda$, then

$$z_i^{r+1} = z_i^{r,i} \leq 0 \quad (4.18)$$

for large enough r .

- If $d_i^\infty = -\lambda$, then

$$z_i^{r+1} = z_i^{r,i} \geq 0 \quad (4.19)$$

for large enough r .

Since $\{f_2(\mathbf{z}^{r,i})\}_{r \geq 0, 1 \leq i \leq n}$ is bounded above, it follows by assumption (A5)* that $\{\mathbf{z}^{r+1}\}_{r \in \mathcal{R}}$ is bounded (with the coordinates in S^c uniformly bounded away from zero), and hence has at least one limit point. Let \mathbf{z}^∞ denote any limit point of $\{\mathbf{z}^{r+1}\}_{r \in \mathcal{R}}$. It follows that

$$E\mathbf{z}^\infty = \mathbf{t}^\infty \text{ and } d(\mathbf{z}^\infty) = 2E^T \nabla \mathbf{t}^\infty = d^\infty. \quad (4.20)$$

It follows by (4.8), (4.16), (4.17), (4.18) and (4.19) that $z_j^\infty = 1/d_j^\infty$ if $j \in S^c$, $z_j^\infty = 0$ if $j \in S$ and $|d_j^\infty| < \lambda$, $z_j^\infty \leq 0$ if $j \in S$ and $d_j^\infty = \lambda$, $z_j^\infty \geq 0$ if $j \in S$ and $d_j^\infty = -\lambda$. It follows from Lemma 4.1 (b) that $\mathbf{z}^\infty \in \mathcal{X}_\ell^*$. It follows by Lemma 3.1 that $\mathbf{t}^\infty = E\mathbf{z}^\infty = \mathbf{t}^*$. The result follows by noting that \mathbf{t}^∞ is an arbitrarily chosen limit point of $\{\mathbf{t}^{r,i}\}_{r \geq 0}$. \square
It follows by Lemma 4.4 and the continuity of g at \mathbf{t}^* that

$$\mathbf{d}(\mathbf{z}^{r,i}) \rightarrow \mathbf{d}^* \quad (4.21)$$

as $r \rightarrow \infty$ for every $1 \leq i \leq n$. The next two lemmas show that the sequence of iterates $\{\mathbf{z}^r\}_{r \geq 0}$ approaches \mathcal{X}_ℓ^* .

Lemma 4.5 *If $i \in S$, then*

$$z_i^r - \text{sign}(z_i^r - d_i(\mathbf{z}^r)) \max(|z_i^r - d_i(\mathbf{z}^r)| - \lambda, 0) \rightarrow 0$$

as $r \rightarrow \infty$. *If $i \in S^c$, then*

$$z_i^r \rightarrow \frac{1}{d_i^*}.$$

as $r \rightarrow \infty$.

The proof of the above lemma is provided in the appendix. As in Section 3, for every $\mathbf{x} \in \mathcal{X}$, define the function ϕ as follows:

$$\phi(\mathbf{x}) = \min_{\mathbf{z}^* \in \mathcal{X}_\ell^*} \|\mathbf{x} - \mathbf{z}^*\|.$$

Hence, $\phi(\mathbf{x})$ is the distance of \mathbf{x} from the closed convex set \mathcal{X}_ℓ^* .

Lemma 4.6

$$\phi(\mathbf{z}^r) \rightarrow 0 \text{ as } r \rightarrow \infty.$$

Proof By (4.2), (4.3), (4.4) and the fact that $E\mathbf{z}^* = \mathbf{t}^*$ for every $\mathbf{z}^* \in \mathcal{X}_\ell^*$, it follows that \mathcal{X}_ℓ^* is the solution set of the linear system of equations given by

$$\begin{aligned} E\mathbf{y} &= \mathbf{t}^*, \mathbf{y} \in \mathcal{X}, \\ y_i &= 0 \text{ if } i \in S, |d_i^*| < \lambda, \\ y_i &\leq 0 \text{ if } i \in S, d_i^* = \lambda, \\ y_i &\geq 0 \text{ if } i \in S, d_i^* = -\lambda, \\ y_i &= \frac{1}{d_i^*} \text{ if } i \in S^c. \end{aligned}$$

Note that the statements of Lemma 3.6 apply exactly to the problem at hand for $i \in S$. Since \mathcal{X}_ℓ^* is non-empty, by Lemma 3.3, for sufficiently large r , there exists $\mathbf{y}^r \in \mathcal{X}_\ell^*$ such that

$$\begin{aligned} \|\mathbf{z}^r - \mathbf{y}^r\| &\leq \theta \left(\|E\mathbf{z}^r - \mathbf{t}^*\| + \sum_{i \in S, d_i^* = \lambda} (z_i^r)^+ + \sum_{i \in S, d_i^* = -\lambda} (-z_i^r)^+ + \sum_{i \in S, |d_i^*| < \lambda} |z_i^r| \right) + \\ &\quad \theta \sum_{i \in S^c} \left| z_i^r - \frac{1}{d_i^*} \right| \\ &= \theta \left(\|E\mathbf{z}^r - \mathbf{t}^*\| + \sum_{i \in S^c} \left| z_i^r - \frac{1}{d_i^*} \right| \right), \end{aligned} \tag{4.22}$$

where θ is a constant only depending on E . The result follows by the definition of ϕ , Lemma 4.4 and Lemma 4.5. \square

Let

$$\begin{aligned} I_1^* &:= \{i \in S : d_i^* = \lambda\}, \\ I_2^* &:= \{i \in S : d_i^* = -\lambda\}, \\ I_3^* &:= \{i \in S : |d_i^*| < \lambda\}. \end{aligned}$$

By Lemma 3.6 (recall that the statements of this lemma apply verbatim for $i \in S$) there exists an $r_0 > 0$ such that

$$\mathbf{z}_{I_3^*}^r = \mathbf{0}, (\mathbf{z}_{I_1^*}^r)^+ = \mathbf{0}, (-\mathbf{z}_{I_2^*}^r)^+ = \mathbf{0} \tag{4.23}$$

for every $r \geq r_0$. Let $M = 2E^T E$. By the assumption that E has no zero column, it follows that $m_{ii} > 0$ for every $1 \leq i \leq n$. As in Section 3, let B denote the lower triangular portion of M , and $C = M - B$ denote the strictly upper triangular portion of M . Since M is a positive semi-definite matrix with strictly positive diagonal entries, it follows that Lemma 3.9 and Lemma 3.10 hold with this choice of M, B and C .

Let $I^* = I_1^* \cup I_2^*$, and

$$\beta = \max_{J \subseteq I^*} \sqrt{|J^c|} \left\{ \left(\frac{\tau_J \|(B_{JJ}^{-1})\| \|M_{JJ}\|}{1 - \rho_J} + \Delta + 1 \right) \|(B_{JJ})^{-1} B_{JJ^c}\| + \frac{\tau_J \|(B_{JJ})^{-1}\| \|M_{JJ}\|}{1 - \rho_J} \right\}.$$

The next lemma is a parallel version of analogous to Lemma 3.11 for the problem at hand.

Lemma 4.7 Consider any $J \subseteq I^*$. If for some two integers $s \geq t \geq r_0$ we have $z_i^r \neq 0$ for every $t+1 \leq r \leq s$ and $i \in J$, then, for any $\mathbf{z}^* \in \mathcal{X}_\ell^*$, there holds

$$\|\mathbf{z}_J^s - \mathbf{z}_J^*\| \leq \Delta \|\mathbf{z}_J^t - \mathbf{z}_J^*\| + \beta \max_{t \leq r \leq s} \|\mathbf{z}_{J^c}^r - \mathbf{z}_{J^c}^*\|_\infty.$$

Proof Since $\Delta \geq 1$, it follows that the claim holds if $s = t$. Suppose $s > t \geq r_0$. Fix any $r \in \{t, \dots, s-1\}$ and $i \in I^*$. Recall that $q(\mathbf{y}) = \mathbf{y}^T \mathbf{y}$. By using exactly the same arguments as in the beginning of the proof of Lemma 3.11, it follows that

$$0 = d_i(\mathbf{z}^{r,i}) - d_i^* = 2E_{\cdot i}^T E \mathbf{z}^{r,i} - 2E_{\cdot i}^T \mathbf{t}^* = E_{\cdot i}^T \nabla^2 q(\mathbf{t}^*)(E \mathbf{z}^{r,i} - \mathbf{t}^*).$$

The result now follows by using exactly the same argument as in the proof of [20, Lemma 9] (starting from [20, Page 12, Line -5] to the end of the proof, replacing E by E^T , and w_J^r by $\mathbf{0}$ throughout). \square

Let $\sigma_0 := 1$ and

$$\sigma_k = \Delta + 3 + \beta + (\beta + 1)\sigma_{k-1} \quad k = 1, 2, \dots, n.$$

It follows from the above definition that $\sigma_k \geq 1$ for every $1 \leq k \leq n$, and is monotonically increasing with k .

Fix $\delta > 0$ arbitrarily. Note that by (4.2), $z_i^* = 1/d_i^*$ for every $i \in S^c$ and every $\mathbf{z}^* \in \mathcal{X}_\ell^*$. By Lemma 4.3 and Lemma 4.6 there exists $r_1 > 0$ such that for every $r \geq r_1$,

$$\phi(\mathbf{z}^r) \leq \delta, \tag{4.24}$$

$$\|\mathbf{z}^{r+1} - \mathbf{z}^r\| \leq \delta, \tag{4.25}$$

$$\|\mathbf{z}_{S^c}^r - \mathbf{z}_{S^c}^*\| \leq \delta, \text{ for every } \mathbf{z}^* \in \mathcal{X}_\ell^*. \tag{4.26}$$

The next three lemmas are parallel versions of Lemma 3.12, Lemma 3.13 and Lemma 3.14 respectively. The proofs of these lemmas follow by repeating the proofs of Lemma 3.12, Lemma 3.13 and Lemma 3.14 verbatim, with the following exceptions: replace n by $|S|$ throughout, replace $i \notin J$ by $i \notin S \setminus J$, and replace μ by 0.

Lemma 4.8 Fix $k \in \{1, 2, \dots, |S|\}$ arbitrarily. If for some nonempty $J \subset I^*$, and some integers $t' > t \geq \max(r_0, r_1)$, we have

$$|z_i^t| > \sigma_k \delta, \quad \forall i \in J, \tag{4.27}$$

$$|z_i^r| \leq \sigma_{k-1} \delta, \quad \forall i \notin J, \forall r = t, t+1, \dots, t'-1, \tag{4.28}$$

then the following hold:

(a) $|z_i^{t'}| > \sigma_{k-1} \delta$ for every $i \in J$.

(b) There exists an $\mathbf{z}^* \in \mathcal{X}_\ell^*$ such that

$$\|\mathbf{z}^r - \mathbf{z}^*\|_\infty \leq \sigma_k \delta, \quad \forall r = t, t+1, \dots, t'-1.$$

Lemma 4.9 Fix $k \in \{1, 2, \dots, |S|\}$ arbitrarily. If for some $J \subseteq I^*$ with $|J| \geq |I^*| - k + 1$ and some interger $t > \max(r_0, r_1)$ we have

$$|z_i^t| > \sigma_k \delta, \quad \forall i \in J, \quad (4.29)$$

$$|z_i^t| \leq \sigma_{k-1} \delta, \quad \forall i \notin S \setminus J, \quad (4.30)$$

then there exists an $\mathbf{z}^* \in \mathcal{X}_\ell^*$ and a $\bar{t} \geq t$ satisfying

$$\|\mathbf{z}^r - \mathbf{z}^*\|_\infty \leq \sigma_k \delta, \quad (4.31)$$

for every $r \geq \bar{t}$.

Lemma 4.10 For any $\delta > 0$, there exists an $\mathbf{z}^* \in \mathcal{X}_\ell^*$ and $\hat{r} > 0$ such that

$$\|\mathbf{z}^r - \mathbf{z}^*\|_\infty \leq \sigma_{|S|} \delta + \delta, \quad (4.32)$$

for every $r \geq \hat{r}$.

We can now prove Theorem 2.2 by repeating the arguments at the end of Section 3 (after the proof of Lemma 3.14) verbatim.

5 Applications

In this section, we demonstrate the utility of Theorem 2.1 and Theorem 2.2. In particular, we use these results to establish convergence of two commonly used cyclic coordinatewise descent algorithms: one arising in high dimensional covariance estimation in the context of graphical models, and another arising in high dimensional logistic regression.

5.1 Convergence of a pseudo likelihood based algorithm for graphical model selection

The CONCORD algorithm, introduced in Khare et al. [15], is a sparse inverse covariance estimation algorithm, which uses cyclic coordinatewise minimization to minimize the function

$$Q_{con}(\Omega) = \sum_{i=1}^p -\log \omega_{ii} + \frac{1}{2} \sum_{i=1}^p \Omega_{\cdot i}^T \hat{\Sigma} \Omega_{\cdot i} + \lambda \sum_{1 \leq i < j \leq p} |\omega_{ij}|, \quad (5.1)$$

subject to the constraint that $\Omega = ((\omega_{ij}))_{1 \leq i, j \leq p}$ is a $p \times p$ symmetric matrix with non-negative diagonal entries. Here $\Omega_{\cdot i}$ denotes the i^{th} column of Ω , p is a fixed positive integer, $\lambda > 0$ is a fixed positive real number, and $\hat{\Sigma}$ is the (observed) sample covariance matrix of n i.i.d. observations from a p -variate distribution. Hence $\hat{\Sigma}$ is positive semi-definite. Let $\Sigma = \Omega^{-1}$ denote the (unknown) true covariance matrix for the underlying p -variate distribution. The CONCORD algorithm provides a sparse estimate of the inverse covariance matrix Ω by minimizing the objective function Q_{con} . Models which induce sparsity in the inverse covariance matrix are known as concentration graphical models, and have gained popularity in statistics, machine learning etc.

As with any sparse covariance estimation algorithm, the CONCORD algorithm is particularly developed to tackle high-dimensional settings, i.e., settings where p is much larger than n . The function $Q_{con}(\Omega)$ is a convex function of Ω , but is not necessarily strictly convex if $n < p$, as the matrix $\hat{\Sigma}$ is singular in this case.

Other pseudo likelihood based sparse inverse covariance estimation algorithms in the literature (see [15] for a list of references) also provide sparse estimates of Ω via cyclic coordinatewise minimization for objective functions which are different from Q_{con} . However, there are no convergence guarantees for the corresponding algorithms. In fact, as shown in [15], it is easy to find (non-pathological) examples where some of these algorithms do not converge. On the other hand, as shown below, the results in this paper can be used to establish convergence of the CONCORD algorithm.

Note that the output produced by the CONCORD algorithm is not guaranteed to be positive definite (same is true for the algorithms in [24, 26]). However, the focus here is to estimate the sparsity pattern in Ω , i.e., model selection. If needed, a positive definite version with the estimated sparsity pattern can be constructed using standard approaches (see Khare et al. [15] for a discussion).

We first provide a lemma which will be useful in our convergence proof.

Lemma 5.1 *Let A be a $k \times k$ positive semi-definite matrix with $A_{kk} > 0$, and λ be a positive constant. Consider the function*

$$h(\mathbf{x}) = -\log x_k + \mathbf{x}^T A \mathbf{x} + \lambda \sum_{i=1}^{k-1} |x_i|$$

defined on $\mathbb{R}^{k-1} \times \mathbb{R}_+$. Then, there exist positive constants a_1 and a_2 (depending only on λ and A), such that

$$h(\mathbf{x}) \geq a_1 x_k - a_2$$

for every $\mathbf{x} \in \mathbb{R}^{k-1} \times \mathbb{R}_+$.

Proof Let $\mathbf{x}_{-k} := (x_i)_{1 \leq i \leq k-1}$, and

$$A = \begin{bmatrix} A_{11} & \mathbf{b} \\ \mathbf{b}^T & A_{kk} \end{bmatrix}.$$

Since A is positive semi-definite, and $A_{kk} > 0$, it follows that

$$\mathbf{x}^T A \mathbf{x} = A_{kk} \left(x_k + \frac{\mathbf{b}^T \mathbf{x}_{-k}}{A_{kk}} \right)^2 + \mathbf{x}_{-k}^T \left(A_{11} - \frac{1}{A_{kk}} \mathbf{b} \mathbf{b}^T \right) \mathbf{x}_{-k} \geq A_{kk} \left(x_k + \frac{\mathbf{b}^T \mathbf{x}_{-k}}{A_{kk}} \right)^2. \quad (5.2)$$

Note that for any $c > 0$, the function $cy - \log y$ is minimized at $y = \frac{1}{c}$. Hence, for every $c > 0$ and $y > 0$, we get that $cy - \log y \geq 1 + \log c$. If $\mathbf{b} = 0$, then it follows by (5.2) and the definition of $h(\mathbf{x})$ that

$$h(\mathbf{x}) \geq -\log x_k + A_{kk} x_k^2 \geq -\log x_k + 2A_{kk} x_k - A_{kk} \geq A_{kk} x_k + 1 + \log A_{kk} - A_{kk}. \quad (5.3)$$

Hence the result holds if $\mathbf{b} = \mathbf{0}$.

If $\mathbf{b} \neq \mathbf{0}$, then $\|\mathbf{b}\|_\infty > 0$. Since $|\mathbf{b}^T \mathbf{x}_{-k}| \leq \|\mathbf{b}\|_\infty \sum_{i=1}^{k-1} |x_i|$, it follows by (5.3) and the definition of $h(\mathbf{x})$ that

$$h(\mathbf{x}) \geq -\log x_k + A_{kk} \left(x_k + \frac{\mathbf{b}^T \mathbf{x}_{-k}}{A_{kk}} \right)^2 + \frac{\lambda A_{kk}}{\|\mathbf{b}\|_\infty} \left| \frac{\mathbf{b}^T \mathbf{x}_{-k}}{A_{kk}} \right|. \quad (5.4)$$

It follows by Lemma 4.1 (b) that for every $x_k > 0$ and $\tilde{\lambda} > 0$, the function $h(y) = (x_k + y)^2 + \tilde{\lambda}|y|$ is minimized at $y = -\left(x_k - \frac{\tilde{\lambda}}{2}\right)_+$. It follows from (5.4) that

$$\begin{aligned} h(\mathbf{x}) &\geq -\log x_k + A_{kk} \left(x_k - \left(x_k - \frac{\lambda}{2\|\mathbf{b}\|_\infty} \right)_+ \right)^2 + \frac{\lambda A_{kk}}{\|\mathbf{b}\|_\infty} \left(x_k - \frac{\lambda}{2\|\mathbf{b}\|_\infty} \right)_+ \\ &\geq -\log x_k + \min \left(A_{kk} x_k^2, \frac{\lambda A_{kk}}{\|\mathbf{b}\|_\infty} \left(x_k - \frac{\lambda}{2\|\mathbf{b}\|_\infty} \right)_+ \right) \end{aligned} \quad (5.5)$$

The result follows from (5.5), the fact that $x_k^2 \geq 2x_k - 1$, and the fact that $cy - \log y \geq 1 + \log c$ (for $c = 1$ and $c = \lambda A_{kk}/(2\|\mathbf{b}\|_\infty)$). \square

The following theorem establishes the convergence of the CONCORD algorithm by using Theorem 2.2.

Theorem 5.1 *If the diagonal entries of $\hat{\Sigma}$ are strictly positive, then the sequence of iterates generated by the cyclic coordinatewise minimization algorithm for Q_{con} converges.*

Proof We will show that the minimization problem in (5.1) is a special case of the minimization problem in (1.2), and satisfies assumption (A5)*. Applying Theorem 2.2 yields the proof of convergence of the CONCORD algorithm.

Let $\mathbf{y} = \mathbf{y}(\Omega) \in \mathbb{R}^{p^2}$ denote a vectorized version of Ω obtained by shifting the corresponding diagonal entry at the bottom of each column of Ω , and then stacking the columns on top of each other. More precisely, if P^i is the $p \times p$ permutation matrix such that $P^i \mathbf{z} = (z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_p, z_i)$ for every $\mathbf{z} \in \mathbb{R}^p$, then

$$\mathbf{y} = \mathbf{y}(\Omega) = ((P^1 \Omega_{\cdot 1})^T, (P^2 \Omega_{\cdot 2})^T, \dots, (P^p \Omega_{\cdot p})^T)^T.$$

Note that since Ω is symmetric, $\omega_{ij} = \omega_{ji}$ for every $1 \leq i < j \leq p$. Let $\mathbf{x} = \mathbf{x}(\Omega) \in \mathbb{R}^{\frac{p(p+1)}{2}}$ be the symmetric version of \mathbf{y} , obtained by removing all ω_{ij} with $i > j$ from \mathbf{y} . More precisely,

$$\mathbf{x} = \mathbf{x}(\Omega) = (\omega_{11}, \omega_{12}, \omega_{22}, \dots, \omega_{1p}, \omega_{2p}, \dots, \omega_{pp})^T.$$

Let \tilde{P} be the $p^2 \times \frac{p(p+1)}{2}$ matrix such that every entry of \tilde{P} is either 0 or 1, exactly one entry in each row of \tilde{P} is equal to 1, and $\mathbf{y} = \tilde{P}\mathbf{x}$. Let \tilde{S} be a $p^2 \times p^2$ block diagonal matrix with p diagonal blocks, and the i^{th} diagonal block is equal to $\tilde{S}^i := \frac{1}{2} P^i \hat{\Sigma} (P^i)^T$. It follows that

$$\begin{aligned} \frac{1}{2} \sum_{i=1}^p \Omega_{\cdot i}^T \hat{\Sigma} \Omega_{\cdot i} &= \frac{1}{2} \sum_{i=1}^p \Omega_{\cdot i}^T (P^i)^T P^i \hat{\Sigma} (P^i)^T P^i \Omega_{\cdot i} = \frac{1}{2} \sum_{i=1}^p (P^i \Omega_{\cdot i})^T (P^i \hat{\Sigma} (P^i)^T) (P^i \Omega_{\cdot i}) \\ &= \mathbf{y}^T \tilde{S} \mathbf{y} \\ &= \mathbf{x}^T \tilde{P}^T \tilde{S} \tilde{P} \mathbf{x}. \end{aligned} \quad (5.6)$$

Note that for every $1 \leq i \leq p$, the matrix $\tilde{S}^i = \frac{1}{2}P^i\hat{\Sigma}(P^i)^T$ is positive semi-definite. Let $\tilde{S}^{1/2}$ denote the $p^2 \times p^2$ block diagonal matrix with p diagonal blocks, such that the i^{th} diagonal block is given by $(\tilde{S}^i)^{1/2}$. Let $E = \tilde{S}^{1/2}\tilde{P}$. It follows by (5.6) that

$$\frac{1}{2} \sum_{i=1}^p \Omega_i^T \hat{\Sigma} \Omega_i = (E\mathbf{x})^T (E\mathbf{x}). \quad (5.7)$$

By the definition of $\mathbf{x}(\Omega)$, we obtain

$$\omega_{ii} = x_{\frac{i(i+1)}{2}} \quad (5.8)$$

for every $1 \leq i \leq p$. Let

$$S = \left\{ j : 1 \leq j \leq \frac{p(p+1)}{2}, j \neq \frac{i(i+1)}{2} \text{ for any } 1 \leq i \leq p \right\},$$

and

$$\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^{\frac{p(p+1)}{2}} : x_j \geq 0 \text{ for every } j \in S^c\}.$$

It follows by (5.1), (5.7) and (5.8) that the CONCORD algorithm can be viewed as a cyclic coordinatewise minimization algorithm to minimize the function

$$Q_{con}(\mathbf{x}) = \mathbf{x}^T E^T E \mathbf{x} - \sum_{i \in S^c} \log x_i + \lambda \sum_{i \in S} |x_i|, \quad (5.9)$$

subject to $\mathbf{x} \in \mathcal{X}$. For any $1 \leq i \leq p(p+1)/2$, there exist $1 \leq k, l \leq p$ such that $x_i = \omega_{kl}$. Note that $\|E_{\cdot i}\|^2 = \frac{\hat{\Sigma}_{kk} + \hat{\Sigma}_{ll}}{2} > 0$. In order to verify assumption (A5)*, we consider the set the set $R_\xi = \{\mathbf{x} : Q_{con}(\mathbf{x}) \leq \xi\}$. Recall by (5.1) that

$$Q_{con}(\mathbf{x}) = Q_{con}(\mathbf{x}(\Omega)) = \sum_{i=1}^p \left\{ -\log \omega_{ii} + \frac{1}{2} \Omega_i^T \hat{\Sigma} \Omega_i + \frac{\lambda}{2} \sum_{1 \leq j \neq i \leq p} |\omega_{ij}| \right\}. \quad (5.10)$$

It follows by applying Lemma 5.1 for every $1 \leq i \leq p$ in (5.10) that there exist positive constants a_1 and a_2 (depending only on $\hat{\Sigma}$ and λ) such that

$$Q_{con}(\mathbf{x}) \geq a_1 \sum_{i=1}^p \omega_{ii} - a_2 = a_1 \sum_{i \in S^c} x_i - a_2. \quad (5.11)$$

Hence, if $\mathbf{x} \in R_\xi$, then

$$x_i \leq (\xi + a_2)/a_1 \leq \tilde{\xi} \quad (5.12)$$

for every $i \in S^c$, where $\tilde{\xi} = (|\xi| + a_2)/a_1$. It also follows by the definition of Q_{con} that if $\mathbf{x} \in R_\xi$, then $-\sum_{i \in S^c} \log x_i < \xi$. Hence $\prod_{i \in S^c} x_i > e^{-\xi}$. It follows by (5.12) that

$$x_i > \frac{e^{-\xi}}{\tilde{\xi}^{p-1}} \quad (5.13)$$

for every $i \in S^c$. It follows by (5.12) that if $\mathbf{x} \in R_\xi$, then

$$\sum_{i \in S} |x_i| \leq \xi + \sum_{i \in S^c} \log x_i \leq \xi + p \log \tilde{\xi}. \quad (5.14)$$

It follows by (5.12), (5.13) and (5.14) that $Q_{con}(\mathbf{x})$ satisfies assumption (A5)*. Combining this with the continuity and convexity of Q_{con} , it follows that the set

$$\mathcal{X}_\ell^* = \{\mathbf{x} \in \mathcal{X} : Q(\mathbf{x}) < \infty, Q(\mathbf{x}^*) \leq Q(\mathbf{x}) \text{ for every } \mathbf{x} \in \mathcal{X}\}.$$

is non-empty. Hence, assumption (A5) holds. It follows by Theorem 2.2 and Remark 1 that the sequence of iterates produced by the CONCORD algorithm converges. \square

Remark Note that if $n \geq 2$, and none of the underlying marginal distributions is degenerate, then the diagonal entries of $\hat{\Sigma}$ are strictly positive, and the assumption in Theorem 5.1 is immediately satisfied.

5.2 Convergence of ℓ_1 minimization for logistic regression

Let Y_1, Y_2, \dots, Y_N denote independent random variables taking values in $\{-1, 1\}$, and $\{\mathbf{z}^i\}_{i=1}^N$ be a collection of vectors in \mathbb{R}^p such that

$$P(Y_i = y \mid \mathbf{z}^i) = \frac{1}{1 + e^{-y_i \boldsymbol{\beta}^T \mathbf{z}^i}}$$

for every $1 \leq i \leq N$. The above statistical model is known as the logistic regression model, and the objective is to estimate the parameter $\boldsymbol{\beta} \in \mathbb{R}^p$. However, in many modern applications, the number of observations N is much less than the number of parameters p . To tackle such a situation, Shevade and Keerthi [28] (see also [12, 17, 18, 23, 35]) propose estimating $\boldsymbol{\beta}$ by minimizing the following objective function:

$$Q_{logit}(\boldsymbol{\beta}) = \sum_{i=1}^N \log \left(1 + e^{-y_i \boldsymbol{\beta}^T \mathbf{z}^i} \right) + \lambda \sum_{j=1}^p |\beta_j|. \quad (5.15)$$

Here y_1, y_2, \dots, y_N denote the observed values of Y_1, Y_2, \dots, Y_N respectively, and $\lambda > 0$ is fixed. The purpose of adding the ℓ_1 penalty term $\lambda \sum_{j=1}^p |\beta_j|$ is to induce sparsity in the parameter estimate. Consider the function $g : \mathbb{R}^N \rightarrow \mathbb{R}$ defined by

$$g(\boldsymbol{\eta}) = \sum_{i=1}^N \log \left(1 + e^{-y_i \eta_i} \right). \quad (5.16)$$

Since

$$\frac{\partial^2}{\partial \eta_i^2} \log \left(1 + e^{-y_i \eta_i} \right) = \frac{y_i^2 e^{-y_i \eta_i}}{(1 + e^{-y_i \eta_i})^2} > 0$$

for every $1 \leq i \leq N$, it follows that g is a strictly convex function. Let X denote the $N \times p$ matrix with i^{th} row given by $(\mathbf{z}^i)^T$. It follows by (5.15) that

$$Q_{logit}(\boldsymbol{\beta}) = g(X\boldsymbol{\beta}) + \lambda \sum_{j=1}^p |\beta_j|. \quad (5.17)$$

Note that in a typical high-dimensional setting, we have $N < p$. Hence, the matrix X is singular, and consequently the function Q_{logit} is *not necessarily strictly convex*.

Shevade and Keerthi [28, Page 2248] propose using cyclic coordinatewise minimization for minimizing Q_{logit} . We note that the final algorithm that they present (see [28, Page 2249]) is a variant where at each iteration, the “best coordinate” is chosen according to an appropriate criterion, and the function is minimized with respect to the chosen coordinate (keeping all the other coordinates fixed). Note that the minimizer with respect to a single coordinate cannot be obtained in closed form in this situation. However, such a minimization involves a convex function on a subset of \mathbb{R} , and numerical methods can be used to obtain the required minimizer accurately in a few steps. In particular, the authors in [28] use a combination of Newton-Raphson and bisection methods. To conclude, coordinatewise minimization (cyclic or the variant approach described above) is a viable approach for this problem, and has been used in applications.

It is claimed in [28] that convergence follows from [2, Prop. 4.1, Chap. 3]. However, the result [2, Prop. 4.1, Chap. 3] states that if $F(\mathbf{x})$ is a convex function, and the sequence $\{\mathbf{x}^t\}_{t \geq 0}$ is generated by using

$$\mathbf{x}^{t+1} = \arg \min \left\{ F(\mathbf{x}) + \frac{1}{2c_t} \|\mathbf{x} - \mathbf{x}^t\|_2^2 \right\},$$

where $\liminf_{t \rightarrow \infty} c_t > 0$; then $\{\mathbf{x}^t\}_{t \geq 0}$ converges to a global minimizer of $F(\mathbf{x})$. Hence, to the best of our understanding, this result is not applicable to coordinatewise minimization in the current setting.

We now show that Theorem 2.1 can be used to provide a proof of convergence of the cyclic coordinatewise minimization algorithm for minimizing Q_{logit} .

Theorem 5.2 *If the matrix X has no zero columns, then the sequence of iterates generated by the cyclic coordinatewise minimization algorithm for Q_{logit} converges.*

Proof Consider the function g defined in (5.16). Note that g is non-negative and $C_g = \mathbb{R}^N$. It follows by (5.17) that the minimization problem for Q_{logit} is a special case of the minimization problem in (1.1) with $m = N$, $n = p$ and $E = X$, and that assumptions (A1)-(A4) are satisfied. Also, if $Q_{\text{logit}}(\boldsymbol{\beta}) \leq \xi$, it follows that $|\beta_j| \leq \xi/\lambda$ for every $1 \leq j \leq p$. Hence, the set $\{\boldsymbol{\beta} : Q_{\text{logit}}(\boldsymbol{\beta}) \leq \xi\}$ is a bounded set for every $\xi \in \mathbb{R}$. It follows that $Q(\boldsymbol{\beta})$ satisfies assumption (A5). The result now follows by Theorem 2.1. \square

Appendix

Proof of Lemma 3.2 First, let us assume that $x \in \mathcal{X}^*$. Then (3.3), (3.4) and (3.5) hold. Hence (3.6) holds automatically. Suppose $i \in S$ and $x_i \neq 0$. By (3.4), it follows that $d_i(\mathbf{x}) + \lambda \text{sign}(x_i) = 0$. Hence,

$$|x_i - d_i(\mathbf{x})| = |x_i + \lambda \text{sign}(x_i)| = |x_i| + \lambda.$$

Since $\lambda, |x_i| > 0$, we obtain $\text{sign}(x_i + \lambda \text{sign}(x_i)) = \text{sign}(x_i)$. It follows that

$$\text{sign}(x_i - d_i(\mathbf{x})) \max(|x_i - d_i(\mathbf{x})| - \lambda, 0) = \text{sign}(x_i) \max(|x_i| + \lambda - \lambda, 0) = x_i.$$

Suppose $i \in S$ and $x_i = 0$. By (3.5), it follows that

$$\text{sign}(x_i - d_i(\mathbf{x})) \max(|x_i - d_i(\mathbf{x})| - \lambda, 0) = \text{sign}(-d_i(\mathbf{x})) \max(|d_i(\mathbf{x})| - \lambda, 0) = 0.$$

This establishes (3.7).

Now, let us assume that (3.6) and (3.7) hold. Hence (3.3) holds automatically. Suppose $i \in S$ and $x_i \neq 0$. We consider three cases.

1. If $x_i = d_i(\mathbf{x})$, then $x_i = 0$ by (3.7), which is a contradiction.
2. If $x_i > d_i(\mathbf{x})$, then $x_i = \max(x_i - d_i(\mathbf{x}) - \lambda, 0)$ by (3.7). Since $x_i \neq 0$, it follows that $x_i > 0$ and $d_i(\mathbf{x}) = -\lambda = -\lambda \text{sign}(x_i)$.
3. If $x_i < d_i(\mathbf{x})$, then $x_i = \min(x_i + \lambda - d_i(\mathbf{x}), 0)$ by (3.7). Since $x_i \neq 0$, it follows that $x_i < 0$ and $d_i(\mathbf{x}) = \lambda = -\lambda \text{sign}(x_i)$.

Hence (3.4) holds. Suppose $i \in S$ and $x_i = 0$. Then $\max(|d_i(\mathbf{x})| - \lambda, 0) = 0$ by (3.7). It follows that $|d_i(\mathbf{x})| \leq \lambda$. Hence, (3.5) holds. It follows that $x \in \mathcal{X}^*$. \square

Proof of Lemma 3.11 Since $\Delta \geq 1$, it follows that the claim holds if $s = t$. Suppose $s > t \geq r_0$. Fix any $r \in \{t, \dots, s-1\}$ and $i \in I^*$. Note that $x_i^{r+1} \neq 0$. If $i \in I_1^*$, it follows from (3.36) that $x_i^{r+1} < 0$. Hence, by (3.8), we obtain $d_i(\mathbf{x}^{r,i}) = \lambda = d_i^*$. If $i \in I_2^*$, it follows from (3.36) that $x_i^{r+1} > 0$. Hence, by (3.8), we obtain $d_i(\mathbf{x}^{r,i}) = -\lambda = d_i^*$. If $i \in I_4^*$, it follows from (3.36) that $x_i^{r+1} > 0$. Hence, by (3.9), we obtain $d_i(\mathbf{x}^{r,i}) = 0 = d_i^*$. In either case, it follows that

$$\begin{aligned} 0 &= d_i(\mathbf{x}^{r,i}) - d_i^* \\ &= d_i(\mathbf{x}^{r,i}) - d_i(\mathbf{x}^*) \\ &= E_{\cdot i}^T (\nabla g(E\mathbf{x}^{r,i}) - g(E\mathbf{x}^*)) \\ &= E_{\cdot i}^T \nabla^2 g(E\mathbf{x}^*) (E\mathbf{x}^{r,i} - E\mathbf{x}^*) + O(\|E\mathbf{x}^{r,i} - E\mathbf{x}^*\|^2). \end{aligned}$$

The result now follows by using exactly the same argument as in the proof of [20, Lemma 9] (starting from [20, Page 12, Line -5] to the end of the proof, after replacing E by E^T throughout). \square

Proof of Lemma 3.12 Let \mathbf{x}^* be any element of \mathcal{X}^* satisfying $\phi(\mathbf{x}^t) = \|\mathbf{x}^t - \mathbf{x}^*\|$. Hence,

$$\|\mathbf{x}^t - \mathbf{x}^*\| \leq \delta. \quad (5.18)$$

By (3.52), if $i \notin J$, then

$$|x_i^*| \leq |x_i^t| + \|\mathbf{x}^t - \mathbf{x}^*\| \leq \sigma_{k-1}\delta + \delta.$$

It follows by (3.52) that

$$||x_i^r| - |x_i^*|| \leq \sigma_{k-1}\delta + \delta,$$

for every $t \leq r \leq t' - 1$. Since $t \geq r_0$ it follows that x_i^r and x_i^* are either both non-positive or both non-negative for every $r \geq t$. Hence $||x_i^r| - |x_i^*|| = |x_i^r - x_i^*|$ for every $r \geq t$, which implies that

$$|x_i^r - x_i^*| \leq \sigma_{k-1}\delta + \delta, \quad (5.19)$$

for every $t \leq r \leq t' - 1$. We now claim that

$$|x_i^r| > \sigma_{k-1}\delta + \delta, \quad (5.20)$$

for every $i \in J$ and $t \leq r \leq t' - 1$. We proceed to prove this by induction. Note that by (3.51) and fact that $\sigma_k \geq \sigma_{k-1} + 1$, it follows that (5.19) holds for $r = t$. Suppose that (5.19) holds for every $t \leq r \leq s$ for some s which satisfies $t \leq s \leq t' - 2$. Hence, $x_r^i \neq 0$ for every $i \in J$ and $t + 1 \leq r \leq s$. It follows by Lemma 3.11 that

$$\|\mathbf{x}_J^s - \mathbf{x}_J^*\| \leq \Delta \|\mathbf{x}_J^t - \mathbf{x}_J^*\| + \beta \max_{t \leq r \leq s} \|\mathbf{x}_{J^c}^r - \mathbf{x}_{J^c}^*\|_\infty + \mu \sum_{r=t}^{s-1} \|\mathbf{x}^r - \mathbf{x}^{r+1}\|^2.$$

By (3.48), (5.18) and (5.19), we obtain

$$\|\mathbf{x}_J^s - \mathbf{x}_J^*\| \leq \Delta\delta + \beta(\sigma_{k-1}\delta + \delta) + \mu\delta. \quad (5.21)$$

Hence, for every $i \in J$, it follows from (3.49), (3.51), (5.18), (5.21), and the definition of σ_k that

$$\begin{aligned} |x_i^{s+1}| &\geq |x_i^t| - \|\mathbf{x}_J^t - \mathbf{x}_J^{s+1}\| \\ &\geq |x_i^t| - (\|\mathbf{x}_J^t - \mathbf{x}_J^*\| + \|\mathbf{x}_J^* - \mathbf{x}_J^s\| + \|\mathbf{x}_J^s - \mathbf{x}_J^{s+1}\|) \\ &> \sigma_k\delta - (\delta + \Delta\delta + \beta\sigma_{k-1}\delta + \beta\delta + \mu\delta + \delta) \\ &= \sigma_{k-1}\delta + \delta. \end{aligned}$$

Thus, by induction, we conclude that (5.20) holds for every $i \in J$ and $t \leq r \leq t' - 1$. It follows by the arguments above that (5.21) holds for every $t \leq s \leq t' - 1$. Note that $\beta > 1$ and $\|\mathbf{y}\|_\infty \leq \|\mathbf{y}\|$ for any vector \mathbf{y} . Hence, by (5.19) and the definition of σ_k , we obtain

$$\|\mathbf{x}^r - \mathbf{x}^*\|_\infty \leq (\Delta + \beta\sigma_{k-1} + \beta + \mu)\delta \leq \sigma_k\delta,$$

for every $t \leq r \leq t' - 1$. This proves part (b) of the required result.

It follows from (3.49) and (5.20) with $r = t' - 1$ that

$$\begin{aligned} |x_i^{t'}| &\geq |x_i^{t'-1}| - \|\mathbf{x}^{t'-1} - \mathbf{x}^{t'}\| \\ &> \sigma_{k-1}\delta + \delta - \delta \\ &= \sigma_{k-1}\delta. \end{aligned}$$

This proves part (a) of the required result. \square

Proof of Lemma 3.13 We proceed by induction on k . If $k = 1$, then $J = I^*$. Hence, by (3.36), $x_i^r = 0$ for every $i \notin J$ and $r \geq t$. By Lemma 3.12 (b), the claim holds for $k = 1$ (the proof of Lemma 3.12 part (b) goes through verbatim even in $t' = \infty$). Suppose now that the result holds for every $1 \leq k \leq h - 1$, for some $h \geq 2$. Choose $J \subseteq I^*$ with $|J| \geq |I^*| - h + 1$ arbitrarily. Let $t > \max(r_0, r_1)$ be such that

$$|x_i^t| > \sigma_k\delta, \quad \forall i \in J, \quad (5.22)$$

$$|x_i^t| \leq \sigma_{k-1}\delta, \quad \forall i \notin J. \quad (5.23)$$

1. **Case 1:** $|x_i^r| \leq \sigma_{h-1}\delta$ for every $i \notin J$ and all $r \geq t$.

Since $|x_i^t| > \sigma_h\delta$ for every $i \in J$, it follows from Lemma 3.12 part (b) that there exists an $\mathbf{x}^* \in \mathcal{X}^*$ such that

$$\|\mathbf{x}^r - \mathbf{x}^*\|_\infty \leq \sigma_h\delta,$$

for every $r \geq t$. Hence the result holds for $k = h$, with $\bar{t} = t$.

2. **Case 2:** There exists an $r > t$ and $i \notin J$ such that $|x_i^r| > \sigma_{h-1}\delta$.

Let t' be the smallest $r > t$ such that $|x_i^r| > \sigma_{h-1}\delta$ for some $i \notin J$. By (5.23), we obtain that $|x_i^r| \leq \sigma_{h-1}\delta$ for every $i \notin J$ and $t \leq r \leq t' - 1$, and by (5.22), $|x_i^t| > \sigma_h\delta$ for every $i \in J$. It follows by Lemma 3.12 part (a) that $|x_i^{t'}| > \sigma_{h-1}\delta$ for every $i \in J$.

Consider the $h + 1$ intervals T_0, T_2, \dots, T_h , where $T_0 = [0, \sigma_0\delta]$, $T_i = (\sigma_{i-1}\delta, \sigma_i\delta]$ for every $1 \leq i \leq h - 1$, and $T_h = (\sigma_{h-1}\delta, \infty)$. Since $|x_i^{t'}| > \sigma_{h-1}\delta$ for some $i \notin J$, it follows that T_h contains at least $|J| + 1$ entries (in absolute value) of the vector $\mathbf{x}^{t'}$. By (3.36) and the fact that $\sigma_0 = 1$, we obtain that at least $n - |I^*|$ entries (in absolute value) of the vector $\mathbf{x}^{t'}$ are contained in J_0 . Note that $|J| \geq |I^*| - h + 1$. Hence, this leaves at most $h - 2$ entries which are contained (in absolute value) in one of the $h - 1$ intervals T_1, \dots, T_{h-1} . By the Pigeon Hole principle, there exists a $q \in \{1, 2, \dots, h - 1\}$ such that $|x_i^{t'}| \notin T_q$ for every $1 \leq i \leq n$. Let h' denote the largest q for which this occurs. Let $J' = \{j : |x_j^{t'}| > \sigma_{h'}\delta\}$. It follows from the observations above that $|J'| \geq |J| + h - h' \geq |J| + 1 - h'$. Note that by (3.36), $J' \subseteq I^*$. Since $h' < h$, the induction hypothesis applied to h' , t' and J' yields the existence of an $\mathbf{x}^* \in \mathcal{X}^*$ and $\bar{t} \geq t'$ such that

$$\|\mathbf{x}^r - \mathbf{x}^*\| \leq \sigma_{h'}\delta,$$

for every $r \geq \bar{t}$. Note that $\sigma_{h'} \leq \sigma_h$. Hence, the result holds for \mathbf{x}^* and $k = h$. This completes the induction on k , and establishes the required result for every $1 \leq k \leq n$. □

Proof of Lemma 3.14 Fix any integer $\bar{r} \geq \max(r_0, r_1)$.

1. **Case 1:** $|x_i^r| \leq \sigma_n\delta$ for every $1 \leq i \leq n$ and $r \geq \bar{r}$.

In this case, let $\mathbf{x}^* \in \mathcal{X}^*$ be such that $\phi(\mathbf{x}^{\bar{r}}) = \|\mathbf{x}^{\bar{r}} - \mathbf{x}^*\|$. It follows by (3.48) that

$$|x_i^*| \leq |x_i^{\bar{r}}| + \|\mathbf{x}^{\bar{r}} - \mathbf{x}^*\| \leq \sigma_n\delta + \delta,$$

for every $1 \leq i \leq n$. Note that $|x_i^r| \leq \sigma_n\delta$ for every $1 \leq i \leq n$ and $r \geq \bar{r}$, and by (3.36), x_i^r and x_i^* are either both non-positive or non-negative. It follows that

$$\|\mathbf{x}^r - \mathbf{x}^*\|_\infty \leq \sigma_n\delta + \delta,$$

for every $r \geq \bar{r}$. Hence, (3.56) holds with $\hat{r} = \bar{r}$.

2. **Case 2:** There exists $t \geq \bar{r}$ and $i \in \{1, 2, \dots, n\}$ such that $|x_i^t| > \sigma_n \delta$.

Let $\tilde{T}_j = (\sigma_{j-1}\delta, \sigma_j\delta]$ for $j = 1, 2, \dots, n$. Note that $|x_i^t|$ does not belong to any of $\tilde{T}_1, \tilde{T}_2, \dots, \tilde{T}_n$. By the Pigeon Hole principle, there exists $q \in \{1, 2, \dots, n\}$ such that T_q does not contain any entry (in absolute value) of the vector \mathbf{x}^t . Let k be the largest q for which this occurs. Let $\tilde{J}' = \{j : |x_j^t| > \sigma_k \delta\}$. Then $|J| \geq n - k + 1$, as $\tilde{T}_{k+1}, \tilde{T}_{k+2}, \dots, \tilde{T}_n$ each contain at least one entry (in absolute value) of \mathbf{x}^t and $|x_i^t| > \sigma_n \delta$. By the definition of \tilde{J}' , it follows that

$$\begin{aligned} |x_j^t| &> \sigma_k \delta, \quad \forall j \in \tilde{J}', \\ |x_j^t| &\leq \sigma_{k-1} \delta, \quad \forall j \notin \tilde{J}'. \end{aligned}$$

It follows by (3.36) and the fact that $\sigma_k \geq 1$ that $\tilde{J}' \subseteq I^*$. Hence, the assumptions of Lemma 3.13 hold with k , \tilde{J}' and t . It follows from Lemma 3.13 that there exists an $\mathbf{x}^* \in \mathcal{X}^*$ and a $\bar{t} \geq t$ such that

$$\|\mathbf{x}^r - \mathbf{x}^*\|_\infty \leq \sigma_k \delta,$$

for every $r \geq \bar{t}$. Since $\sigma_k \leq \sigma_n$, we conclude that (3.56) holds with \mathbf{x}^* and $\hat{r} = \bar{t}$. □

Proof of Lemma 4.5 The result for $i \in S$ follows by repeating exactly the same arguments as in the proof of Lemma 3.7 (b) for this case. The result for $i \in S^c$ follows by (4.8) and (4.21). □

References

- [1] Auslender, A. (1978). Minimisation de fonctions localement lipschitziennes: applications a la programmation mi-convexe, mi-differentiable. In: *Mangasarian, O.L., Meyer, R.R., and Robinson, S.M. (eds.) Nonlinear Programming* **3**, pp. 429-460. Academic, New York.
- [2] Bertsekas, D.P. and Tsitsiklis, J.N. (1989). *Parallel and Distributed Computation: Numerical Methods*, Prentice Hall, Englewood Cliffs, NJ, USA.
- [3] Bradley, P.S., Fayyad, U.M. and Mangasarian, O.L. (1999). Mathematical programming for data mining: formulations and challenges, *INFORMS J. Comput.* **11**, 217-238.
- [4] Burke, J.V. (1985). Descent methods for composite nondifferentiable optimization problems, *Math. Program.* **33**, 260-279.
- [5] Chen, S., Donoho, D. and Saunders, M. (1999). Atomic decomposition by basis pursuit, *SIAM J. Sci. Comput.* **20**, 3361.
- [6] Donoho, D.L. and Johnstone, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage, *Biometrika* **81**, 425-455.

- [7] Donoho, D.L. and Johnstone, I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage, *J. Am. Stat. Assoc.* **90**, 12001224.
- [8] Fletcher, R. (1982). A model algorithm for composite nondifferentiable optimization problems. *Math. Program. Study* **17**, 6776.
- [9] Friedman, J., Hastie, T. and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso, *Biostatistics* **9**, 432441.
- [10] Fu, W. (1998). Penalized regressions: the bridge vs the lasso, *Journal of Computational and Graphical Statistics* **3**, 397-416.
- [11] Fukushima, M. and Mine, H. (1981). A generalized proximal point algorithm for certain non-convex minimization problems, *Int. J. Syst. Sci.* **12**, 9891000.
- [12] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software* **33**.
- [13] Friedman, J., Hastie, T., Hofling, H. and Tibshirani, R. (2007). Pathwise coordinatewise optimization, *Annals of Applied Statistics* **1**, 302-332.
- [14] Hoffman, A.J. (1952). On Approximate solutions of systems of linear inequalities, *Journal of Research of the National Bureau of Standards* **49**, 263-265.
- [15] Khare, K., Oh, S. and Rajaratnam, B. (2014). A convex pseudo-likelihood framework for high dimensional partial correlation estimation with convergence guarantees, to appear in *Journal of the Royal Statistical Society B*.
- [16] Kiwiel, K.C. (1986). A method for minimizing the sum of a convex function and a continuously differentiable function, *J. Optim. Theory Appl.* **48**, 437449.
- [17] Koh, K., Kim, S.-J. and Boyd, S. (2007). An interior-point method for large-scale ℓ_1 -regularized logistic regression, *J. Mach. Learn. Res.* **8**, 15191555.
- [18] Lee, S., Lee, H., Abeel, P. and Ng, A. (2006) Efficient l1-regularized logistic regression, In *Proceedings of the 21st National Conference on Artificial Intelligence, 2006*.
- [19] Luo, Z. and Tseng, P. (1989) On the convergence of a matrix splitting algorithm for the symmetric linear complementarity problem, Technical Report LIDS-P 1884, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology.
- [20] Luo, Z. and Tseng, P. (1989) On the convergence of the coordinate descent method for convex differentiable minimization, Technical Report LIDS-P 1924, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology.
- [21] Luo, Z. and Tseng, P. (1992) On the convergence of the coordinate descent method for convex differentiable minimization, *Journal of Optimization Theory and Applications* **72**, 7-35.

- [22] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso, *The Annals of Statistics* **34**, 1436-1462.
- [23] Park, M. and Hastie, T. (2007). L_1 -regularization path algorithm for generalized linear models, *Journal of the Royal Statistical Society B* **69**, 659-677.
- [24] Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models, *Journal of the American Statistical Association* **104**, 735-746.
- [25] Richtarik, P. and Takac, M. (2011). Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function, available on *Arxiv*.
- [26] Rocha, G., Zhao, P., and Yu, B. (2008). A path following algorithm for sparse pseudo-likelihood inverse covariance estimation (SPLICE). Technical report, Statistics Department, UC Berkeley, Berkeley, CA.
- [27] Saha, A. and Tewari, A. (2010). On the Finite Time Convergence of Cyclic Coordinate Descent Methods, *CoRR* abs/1005.2146.
- [28] Shevade, S.K. and Keerthi, S.S. (2003). A simple and efficient algorithm for gene selection using sparse logistic regression, *Bioinformatics* **19**, 2246-2253.
- [29] Sardy, S., Bruce, A. and Tseng, P. (2001). Robust wavelet denoising, *IEEE Trans. Signal Proc.* **49**, 1146-1152.
- [30] Sardy, S. and Tseng, P. (2004). AMlet, RAMlet, and GAMlet: automatic nonlinear fitting of additive models, robust and generalized, with wavelets, *J. Comput. Graph. Stat.* **13**, 283-309.
- [31] Tibshirani, R. (1995). Regression selection and shrinkage via the lasso, *J. Royal Statist. Soc. B* **57**, pp. 267-288.
- [32] Tibshirani, R., Saunders, M., Rosset, S. and Knight, K. (2005). Sparsity and smoothness via the fused lasso, *J. Royal Statist. Soc. B* **67**, 91-108.
- [33] Tseng, P. (2001). Convergence of block coordinate descent method for nondifferentiable minimization, *J. Optim. Theory Appl.* **109**, 473-492.
- [34] Tseng, P. and Yun, S. (2009). A coordinate gradient descent method for nonsmooth separable minimization, *Math. Program., Ser. B* **117**, 387-423.
- [35] Yun, S. and Toh, K. (2011). A coordinate gradient descent method for ℓ_1 -regularized convex minimization, *Computational Optimization and Applications* **48**, 273-307.